Office

WO 99/56223

PATENTS STATUS INFORMATION

FULL DETAILS

REGISTER ENTRY FOR GB2336698

Form 1 Application No GB9808805.7 filing date 24.04.1998

Title ASSOCIATING FILES OF DATA

Applicant/Proprietor
THE DIALOG CORPORATION PLC, Incorporated in the United Kingdom, The Communications Building, 48 Leicester Square, LONDON, WC2H 7DB, United Kingdom [ADP No. 07399819001]

Inventor
MARCUS ALEXANDER BAGSHAW, 69 Victoria Road, CHELMSFORD, Essex, CM1 1PA, United Kingdom [ADP No. 07450778001]

Classified to
G4A
G06F

Address for Service
ATKINSON & CO, PO Box 1205, SHEFFIELD, S9 3UR, United Kingdom [ADP No. 07306061001]

Publication No GB2336698 dated 27.10.1999

Examination requested 06.10.1999

02.09.1998 Notification of change of Address For Service address of
ATKINSON & CO, PO Box 1205, SHEFFIELD, S9 3UR, United Kingdom [ADP No. 07306061001]
to
ATKINSON & CO, First Floor,Unit A, The Technology Park,60 Shirland Lane, SHEFFIELD, S9 3PA, United Kingdom [ADP No. 06477723003]
dated 21.08.1998. Written notification filed on GY19

24.01.2000 Notification of change of Address For Service name and address of
ATKINSON & CO, First Floor,Unit A, The Technology Park,60 Shirland Lane, SHEFFIELD, S9 3PA, United Kingdom [ADP No. 06477723003]
to
ATKINSON BURRINGTON, The Technology Park, 60 Shirland Lane, SHEFFIELD, S9 3PA, United Kingdom [ADP No. 07807043001]
dated 29.12.1999. Official evidence filed on GB2294084

**** END OF REGISTER ENTRY ****

New Enquiry

≡WO 99 56223

(12) UK Patent Application (19) GB (11) 2 336 698 (13) A

(43) Date of A Publication 27.10.1999

(21) Application No 9808805.7

(22) Date of Filing 24.04.1998

(51) INT CL[6]
G06F 17/30

(52) UK CL (Edition Q)
G4A AUDB

(56) Documents Cited
None

(58) Field of Search
UK CL (Edition P) G4A AUDB
INT CL[6] G06F 17/30
Online: LISA

(71) Applicant(s)
The Dialog Corporation Plc
(Incorporated in the United Kingdom)
The Communications Building, 48 Leicester Square,
LONDON, WC2H 7DB, United Kingdom

(72) Inventor(s)
Marcus Alexander Bagshaw

(74) Agent and/or Address for Service
Atkinson & Co
First Floor, Unit A, The Technology Park,
60 Shirland Lane, SHEFFIELD, S9 3PA,
United Kingdom

(54) Abstract Title
Automatic content categorisation of text data files using subdivision to reduce false classification

(57) Text data files are associated with categories by processing the text files with outline files. Large files 221 are subdivided 222 into file sections 223 of comparable length consistent with a preferred size. Each file section is categorised 224 and the original undivided file is assigned 226 to a category only if a sufficient proportion of its constituent sections have been so classified. This technique reduces false classifications.

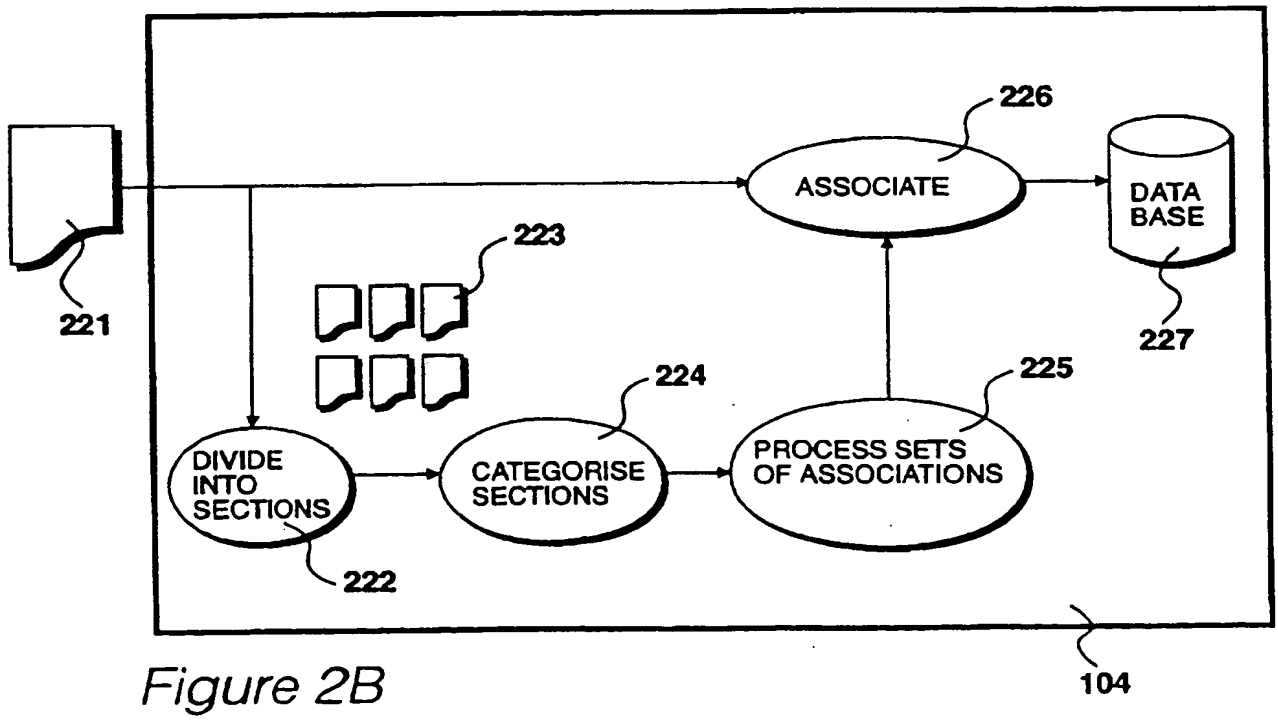

Figure 2B

GB 2 336 698 A

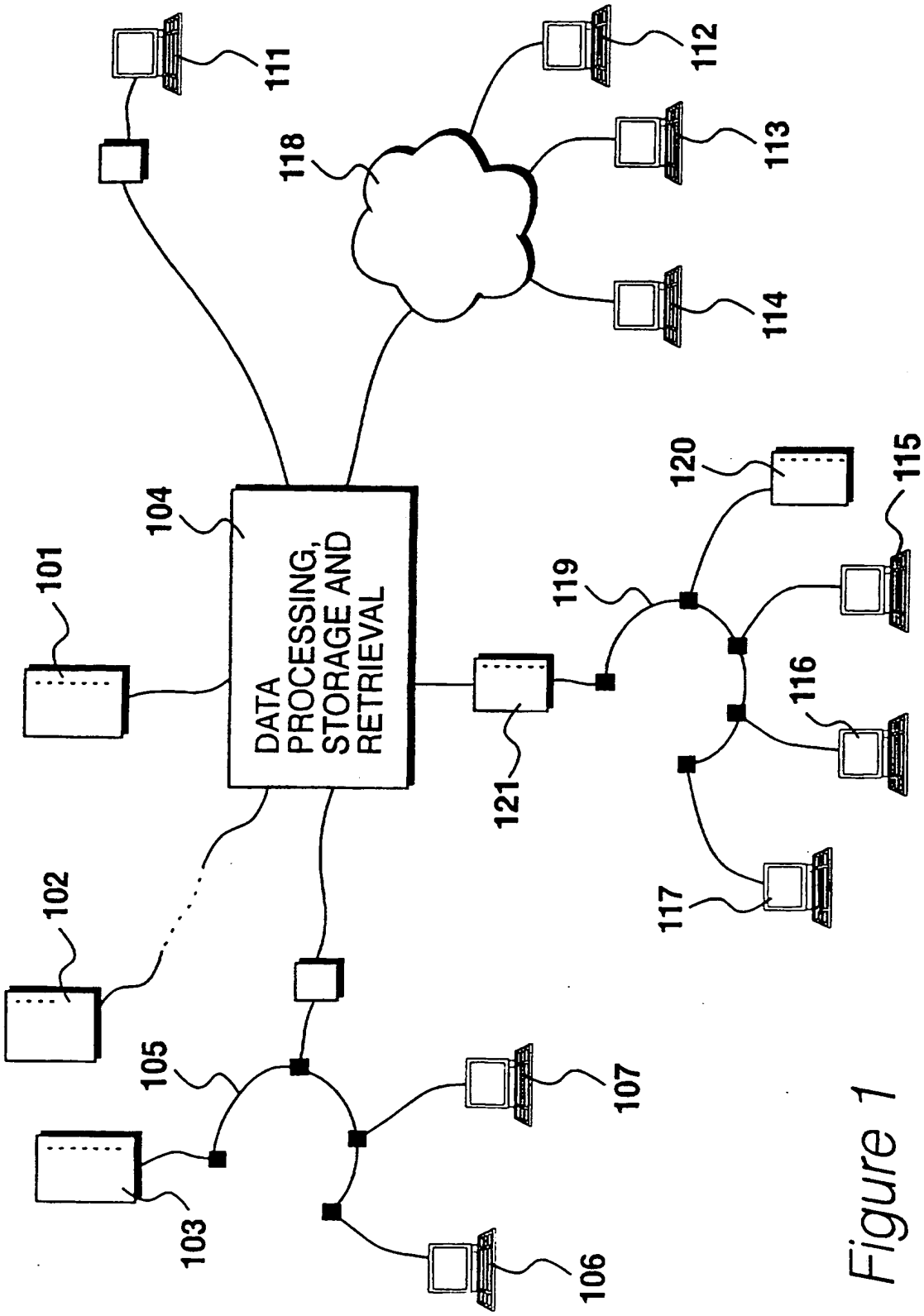
DATA PROCESSING, STORAGE AND RETRIEVAL
101
102
103
104
105
106
107
111
112
113
114
115
116
117
118
119
120
121

Figure 1

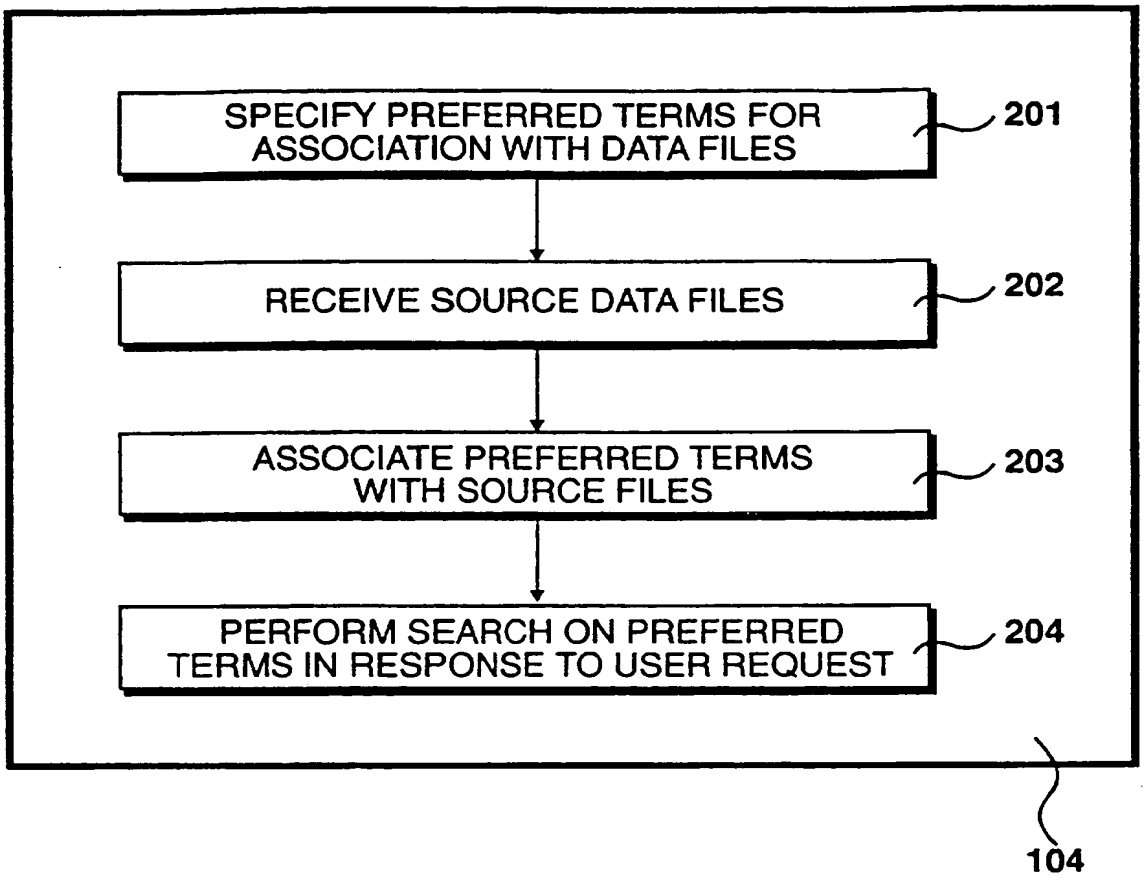
SPECIFY PREFERRED TERMS FOR ASSOCIATION WITH DATA FILES
201
RECEIVE SOURCE DATA FILES
202
ASSOCIATE PREFERRED TERMS WITH SOURCE FILES
203
PERFORM SEARCH ON PREFERRED TERMS IN RESPONSE TO USER REQUEST
204
104

Figure 2A

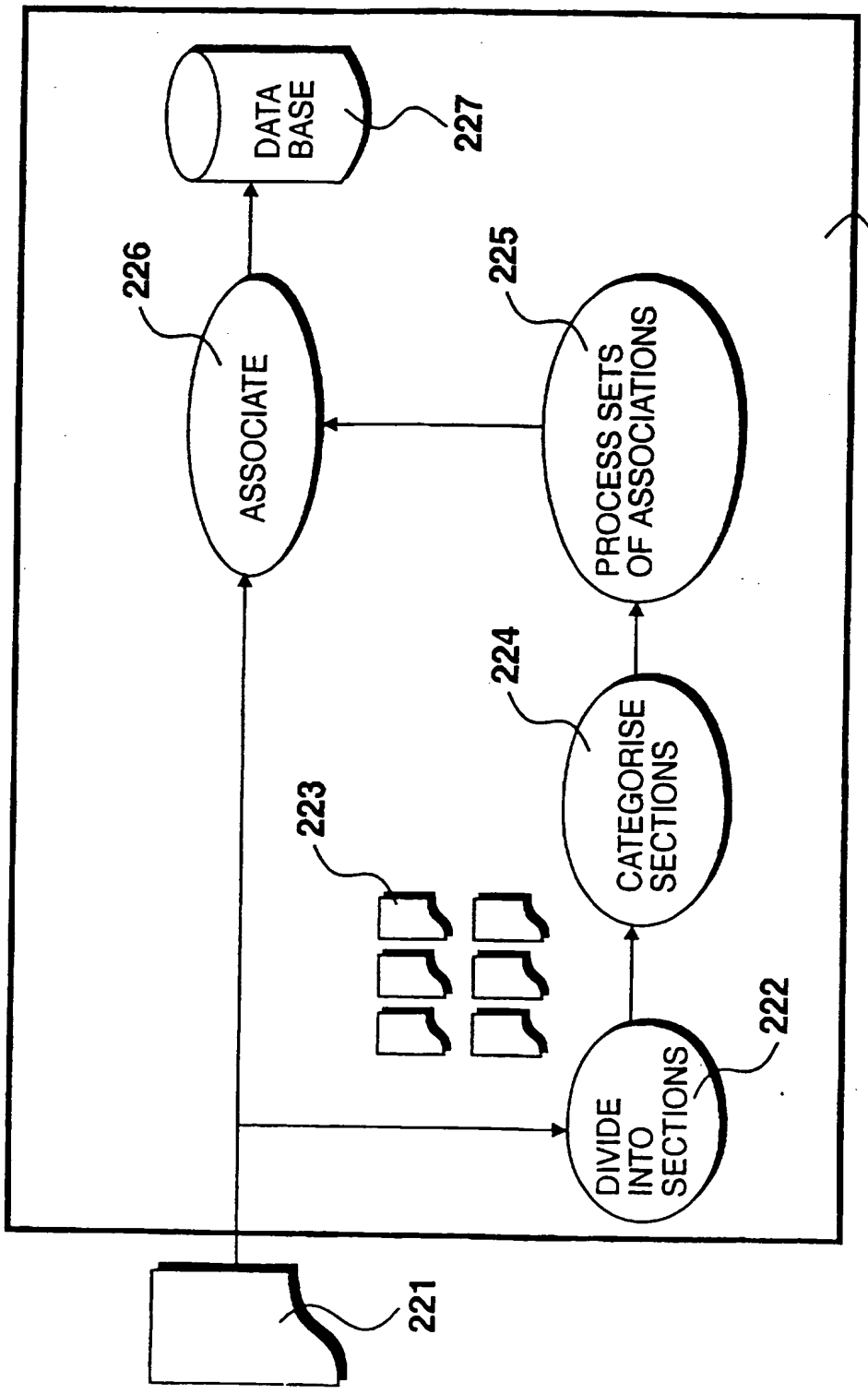
DATA BASE
227
104
226
ASSOCIATE
225
PROCESS SETS OF ASSOCIATIONS
224
CATEGORISE SECTIONS
223
DIVIDE INTO SECTIONS
222
221

Figure 2B

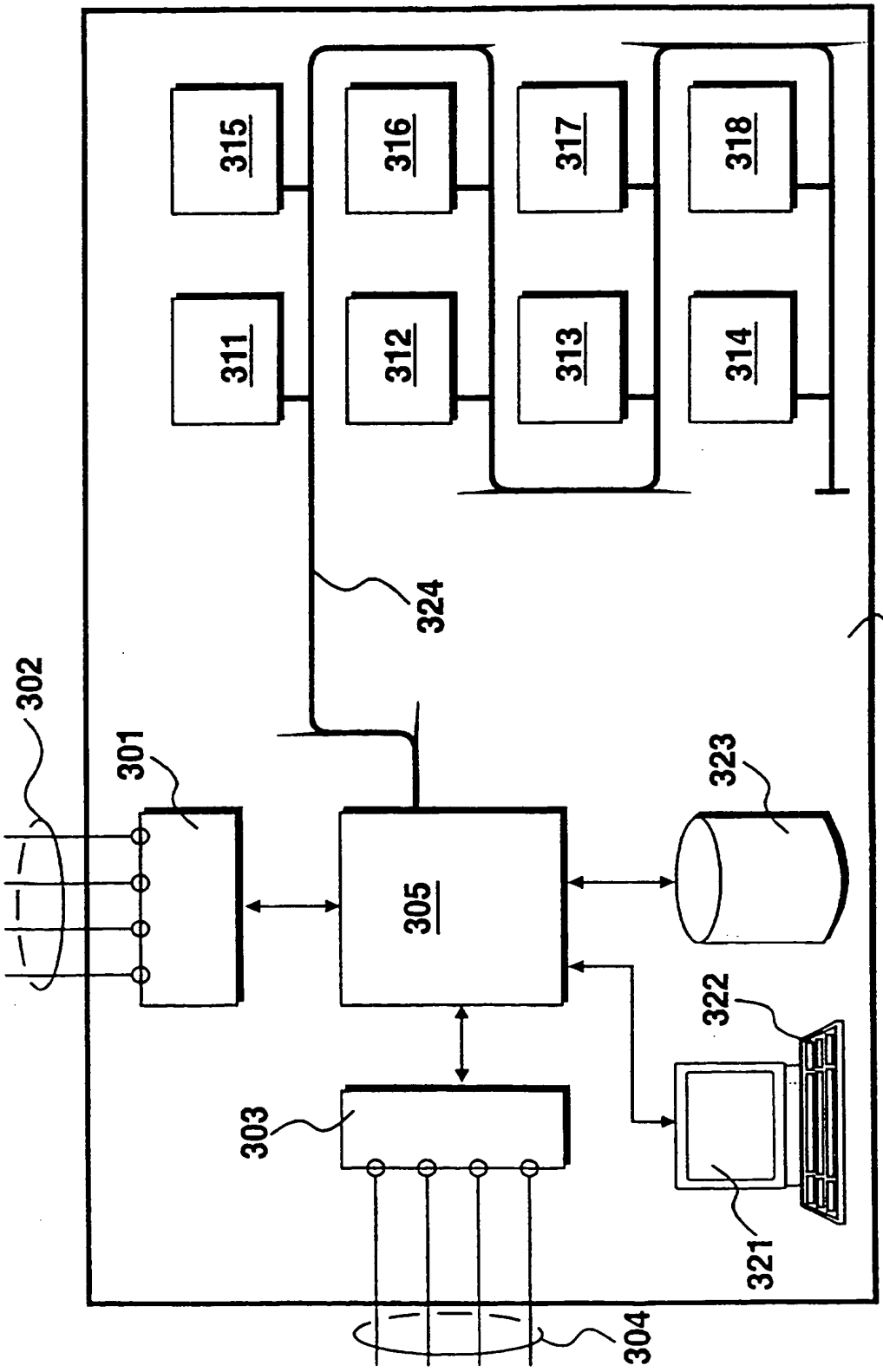
315
316
317
318
311
312
313
314
324
104
302
301
323
305
322
303
321
304

Figure 3

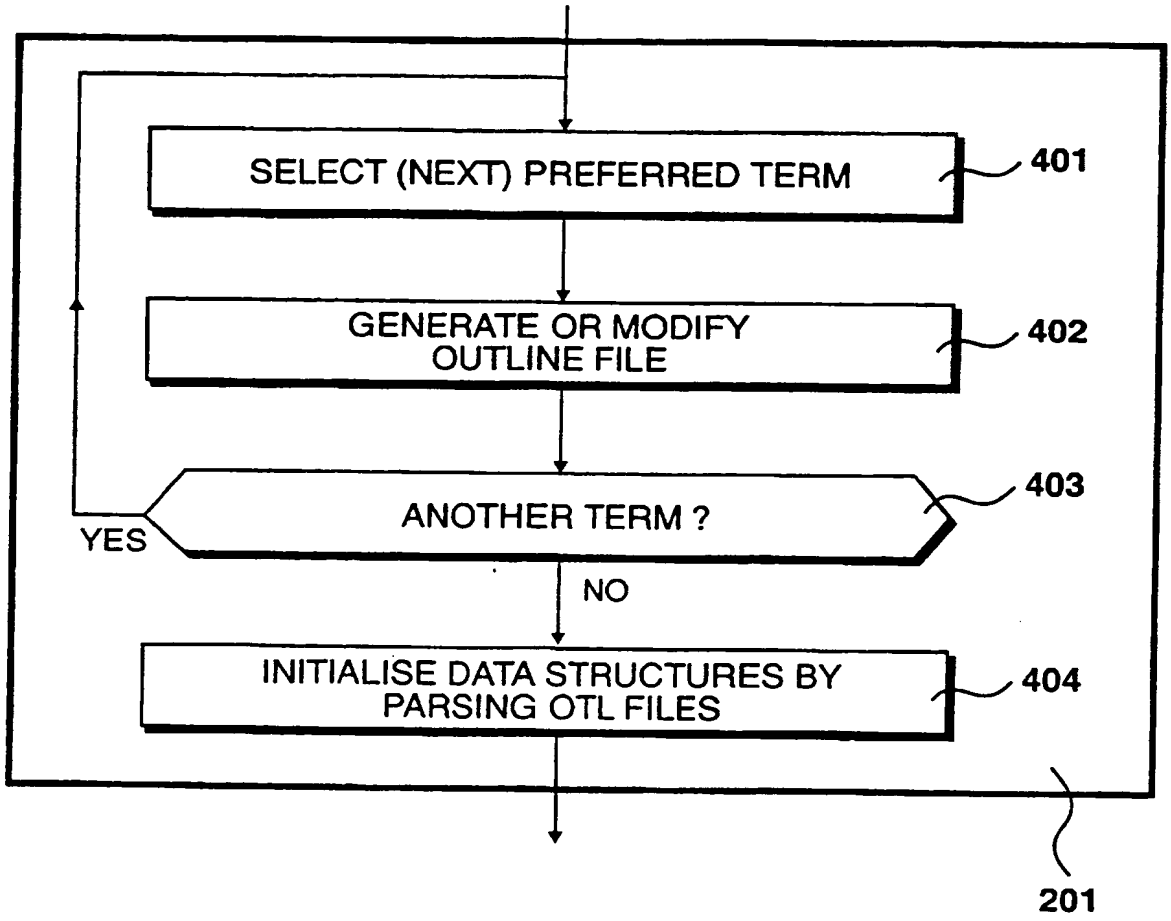
SELECT (NEXT) PREFERRED TERM
401
GENERATE OR MODIFY OUTLINE FILE
402
ANOTHER TERM ?
403
YES
NO
INITIALISE DATA STRUCTURES BY PARSING OTL FILES
404
201

Figure 4

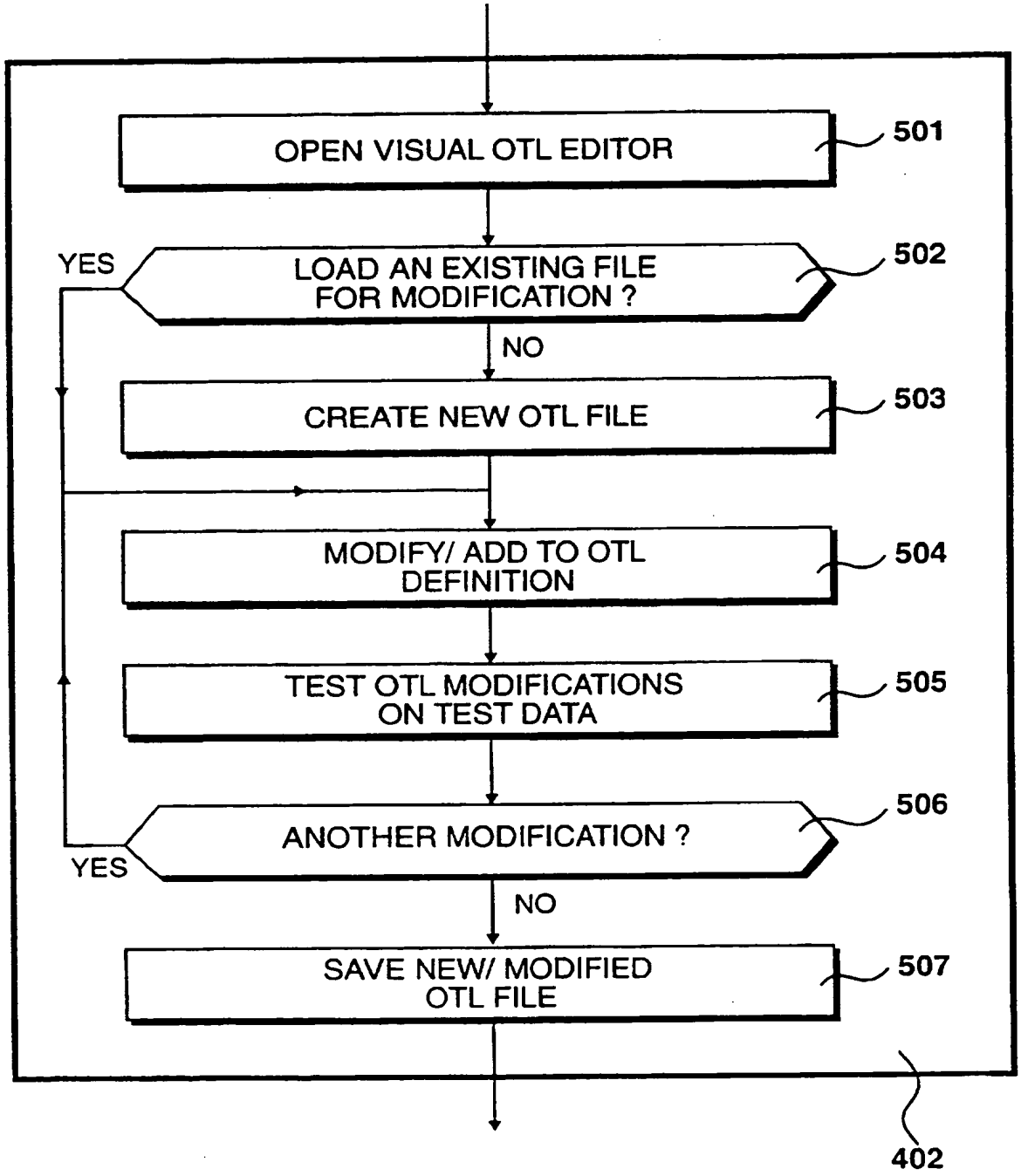
OPEN VISUAL OTL EDITOR
501
YES
LOAD AN EXISTING FILE FOR MODIFICATION ?
502
NO
CREATE NEW OTL FILE
503
MODIFY/ ADD TO OTL DEFINITION
504
TEST OTL MODIFICATIONS ON TEST DATA
505
ANOTHER MODIFICATION ?
506
YES
NO
SAVE NEW/ MODIFIED OTL FILE
507
402

Figure 5

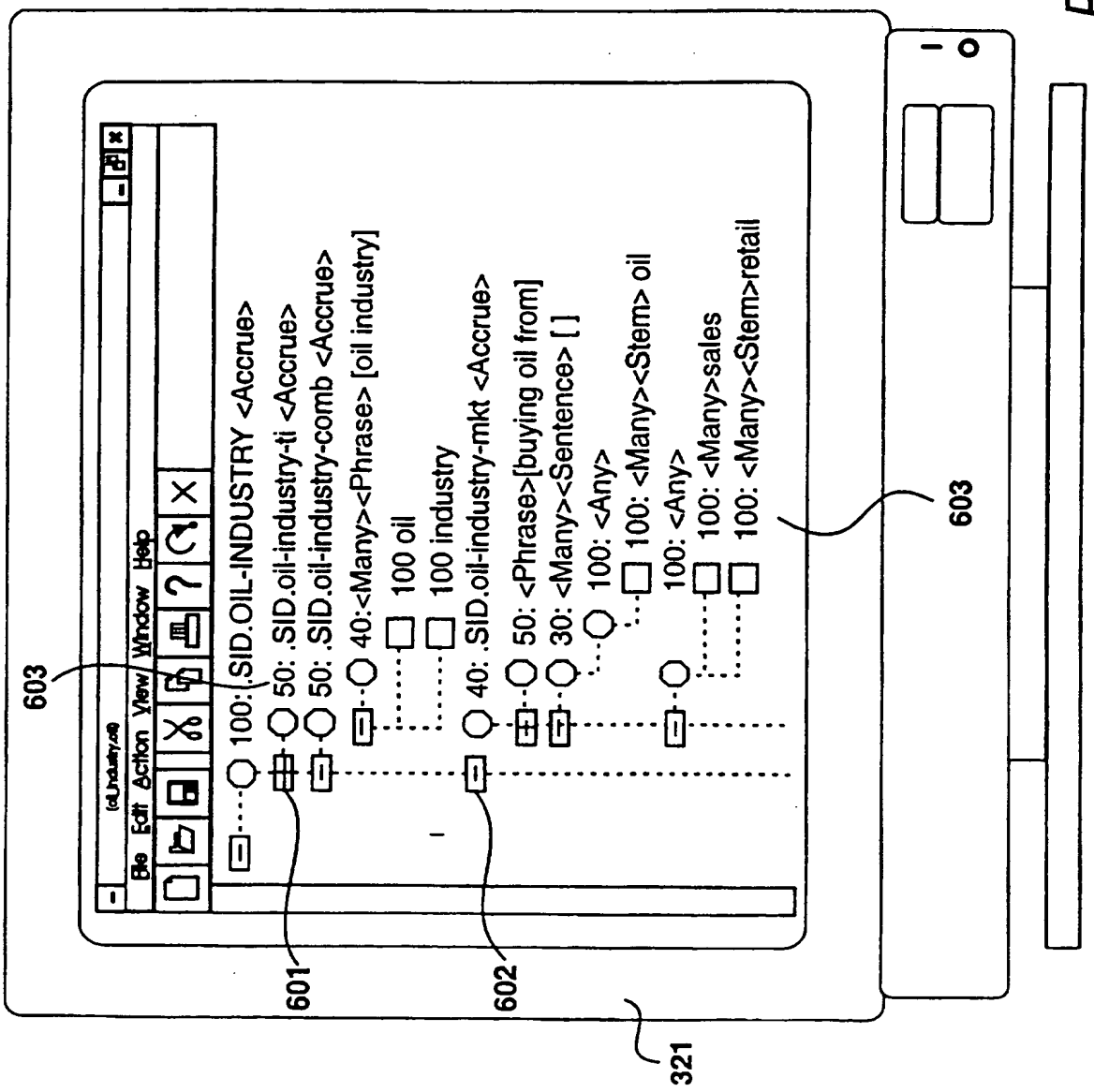
File Edit Action View Window Help
100: SID.OIL-INDUSTRY <Accrue>
50: .SID.oil-industry-ti <Accrue>
50: .SID.oil-industry-comb <Accrue>
40:<Many><Phrase> [oil industry]
100 oil
100 industry
40: .SID.oil-industry-mkt <Accrue>
50: <Phrase>[buying oil from]
30: <Many><Sentence> [ ]
100: <Any>
100: <Many><Stem> oil
100: <Any>
100: <Many>sales
100: <Many><Stem>retail
603
603
601
602
321

Figure 6

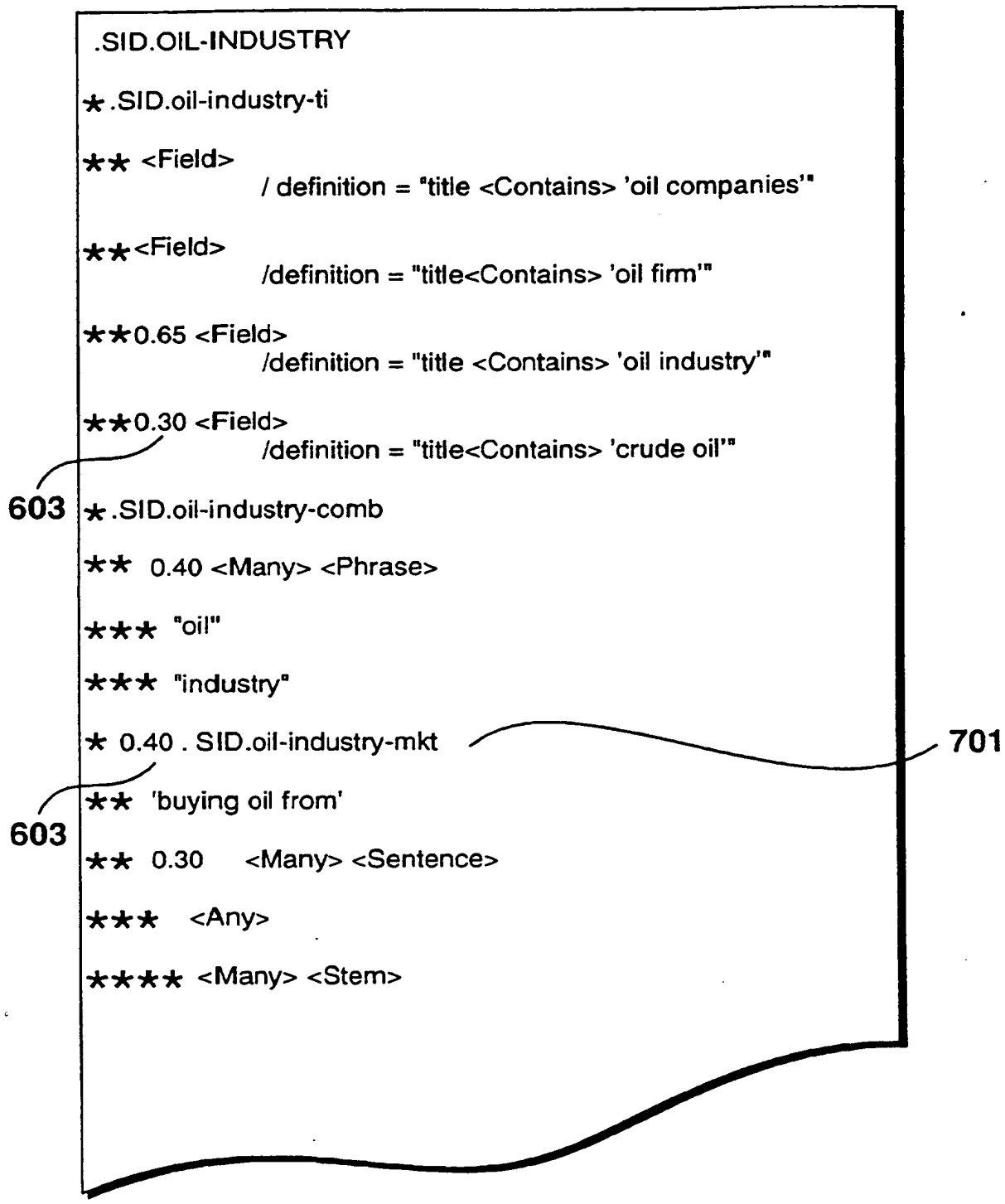
.SID.OIL-INDUSTRY
*.SID.oil-industry-ti
** <Field>
/ definition = "title <Contains> 'oil companies'"
**<Field>
/definition = "title<Contains> 'oil firm'"
**0.65 <Field>
/definition = "title <Contains> 'oil industry'"
**0.30 <Field>
/definition = "title<Contains> 'crude oil'"
603
*.SID.oil-industry-comb
** 0.40 <Many> <Phrase>
*** "oil"
*** "industry"
* 0.40 . SID.oil-industry-mkt
701
** 'buying oil from'
603
** 0.30 <Many> <Sentence>
*** <Any>
**** <Many> <Stem>

Figure 7

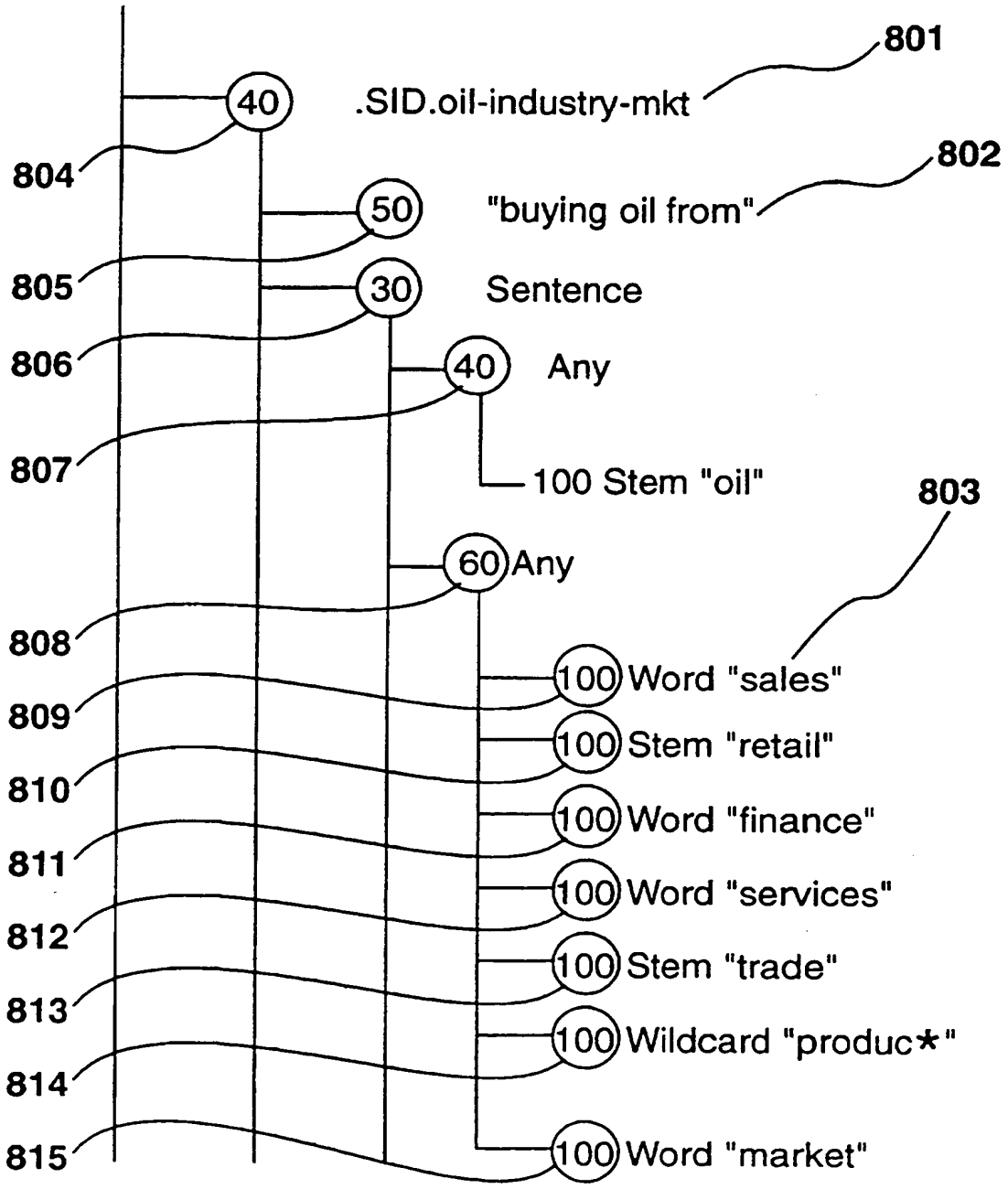
801
40
.SID.oil-industry-mkt
804
802
50
"buying oil from"
805
30
Sentence
806
40
Any
807
100 Stem "oil"
803
60
Any
808
100
Word "sales"
809
100
Stem "retail"
810
100
Word "finance"
811
100
Word "services"
812
100
Stem "trade"
813
100
Wildcard "produc*"
814
815
100
Word "market"

Figure 8

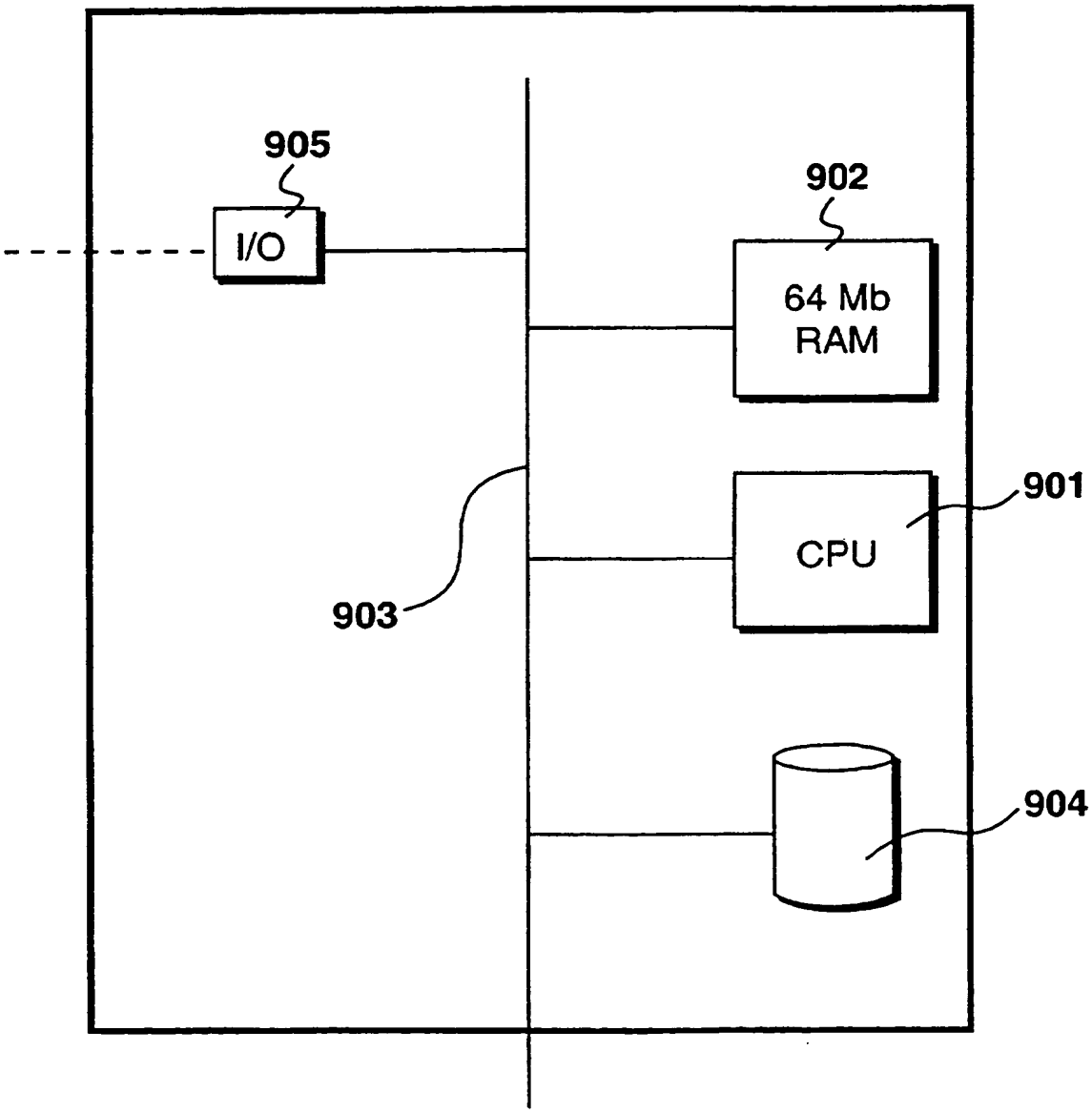

Figure 9

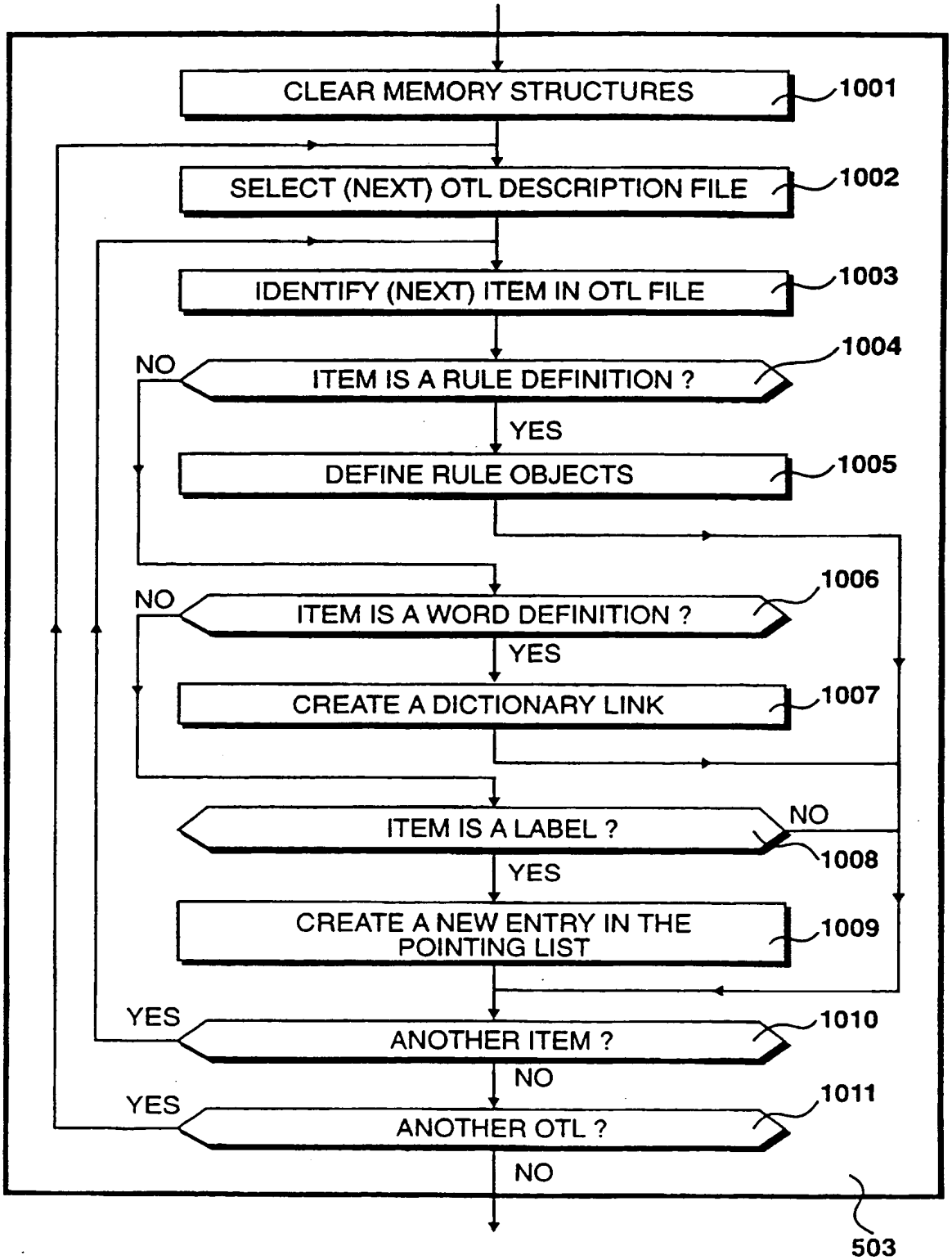
CLEAR MEMORY STRUCTURES
1001
SELECT (NEXT) OTL DESCRIPTION FILE
1002
IDENTIFY (NEXT) ITEM IN OTL FILE
1003
NO
ITEM IS A RULE DEFINITION ?
1004
YES
DEFINE RULE OBJECTS
1005
NO
ITEM IS A WORD DEFINITION ?
1006
YES
CREATE A DICTIONARY LINK
1007
ITEM IS A LABEL ?
NO
1008
YES
CREATE A NEW ENTRY IN THE POINTING LIST
1009
YES
ANOTHER ITEM ?
1010
NO
YES
ANOTHER OTL ?
1011
NO
503

Figure 10

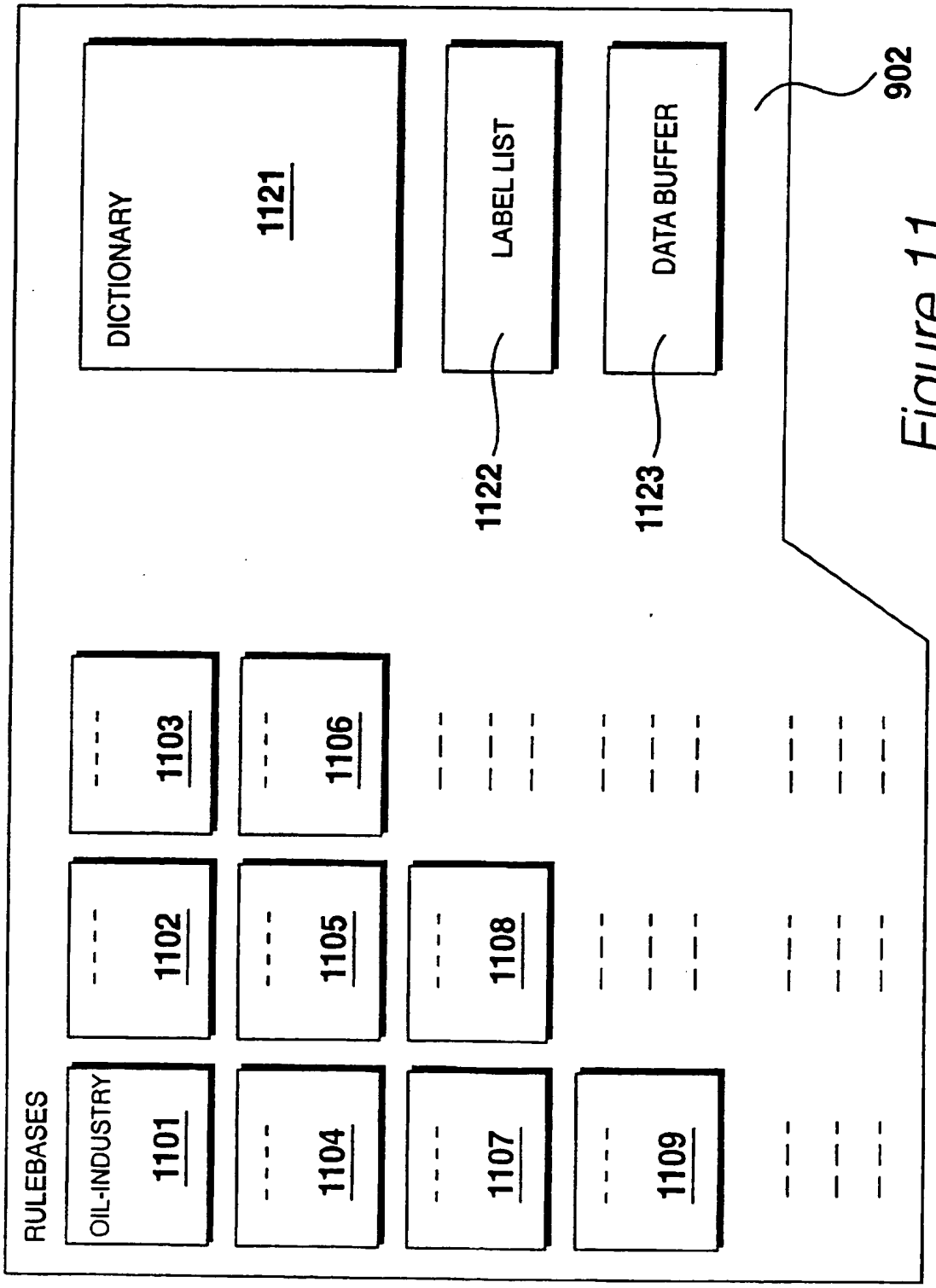
RULEBASES
OIL-INDUSTRY
1101
1102
1103
1104
1105
1106
1107
1108
1109
DICTIONARY
1121
LABEL LIST
1122
DATA BUFFER
1123
902

Figure 11

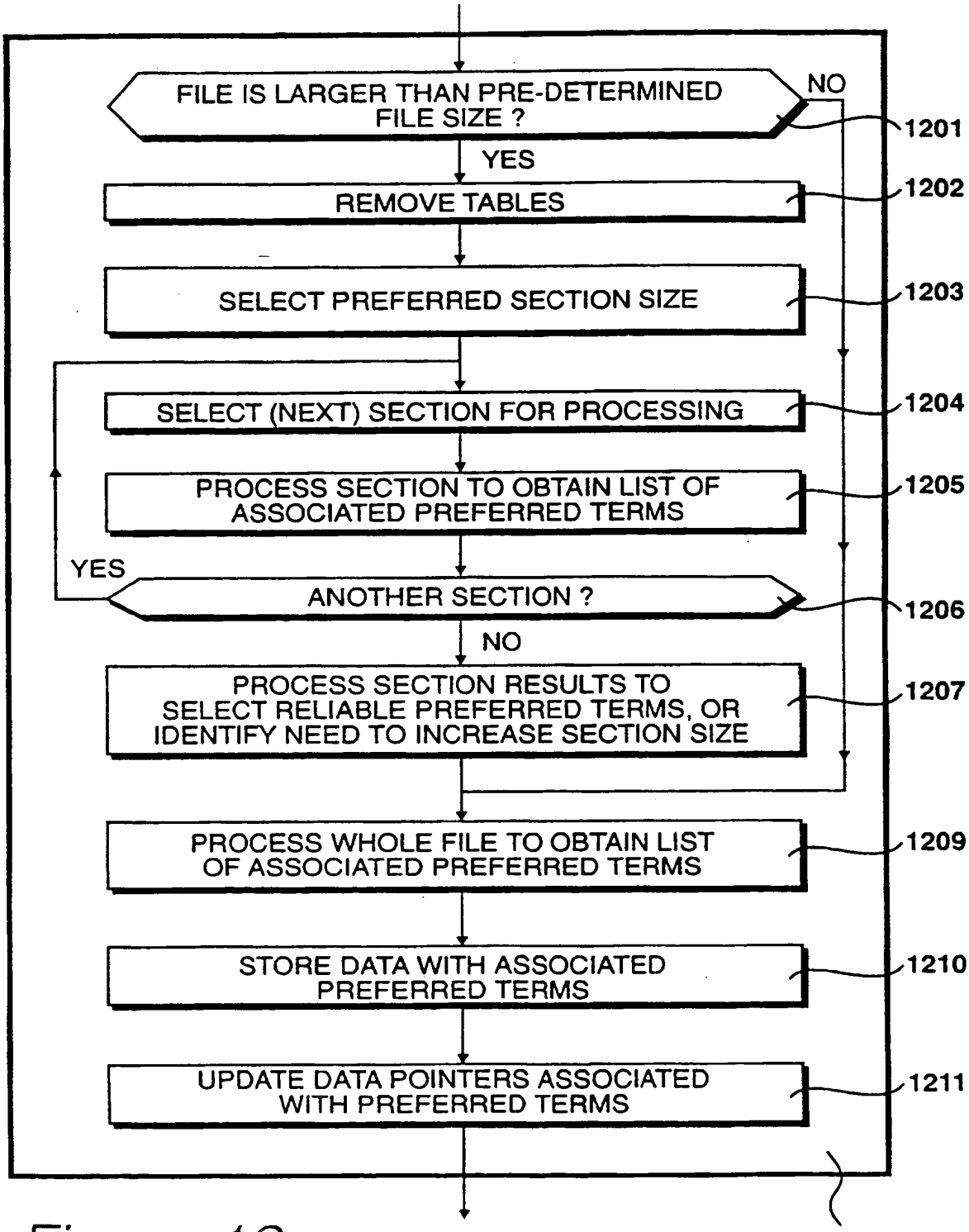
FILE IS LARGER THAN PRE-DETERMINED FILE SIZE ?
NO
1201
YES
REMOVE TABLES
1202
SELECT PREFERRED SECTION SIZE
1203
SELECT (NEXT) SECTION FOR PROCESSING
1204
PROCESS SECTION TO OBTAIN LIST OF ASSOCIATED PREFERRED TERMS
1205
YES
ANOTHER SECTION ?
1206
NO
PROCESS SECTION RESULTS TO SELECT RELIABLE PREFERRED TERMS, OR IDENTIFY NEED TO INCREASE SECTION SIZE
1207
PROCESS WHOLE FILE TO OBTAIN LIST OF ASSOCIATED PREFERRED TERMS
1209
STORE DATA WITH ASSOCIATED PREFERRED TERMS
1210
UPDATE DATA POINTERS ASSOCIATED WITH PREFERRED TERMS
1211
203

Figure 12

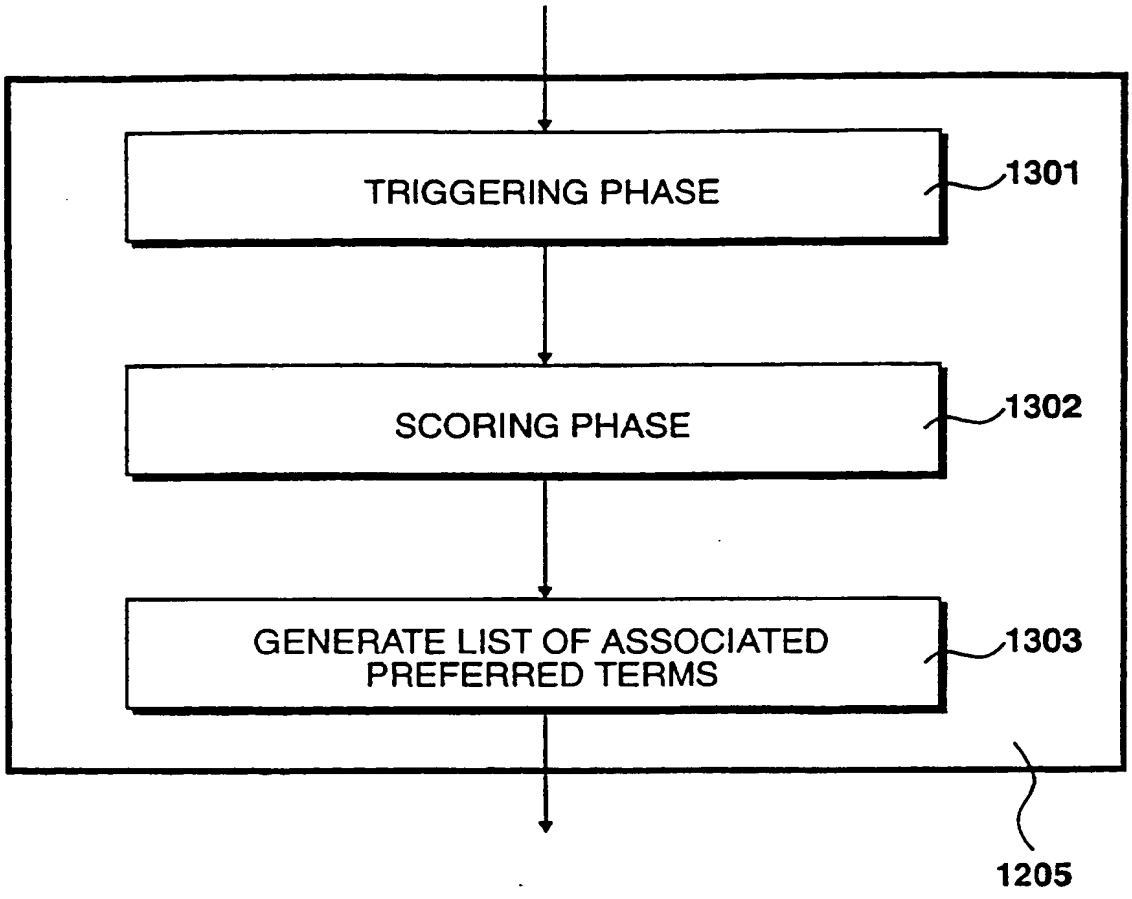
TRIGGERING PHASE
1301
SCORING PHASE
1302
GENERATE LIST OF ASSOCIATED PREFERRED TERMS
1303
1205

Figure 13

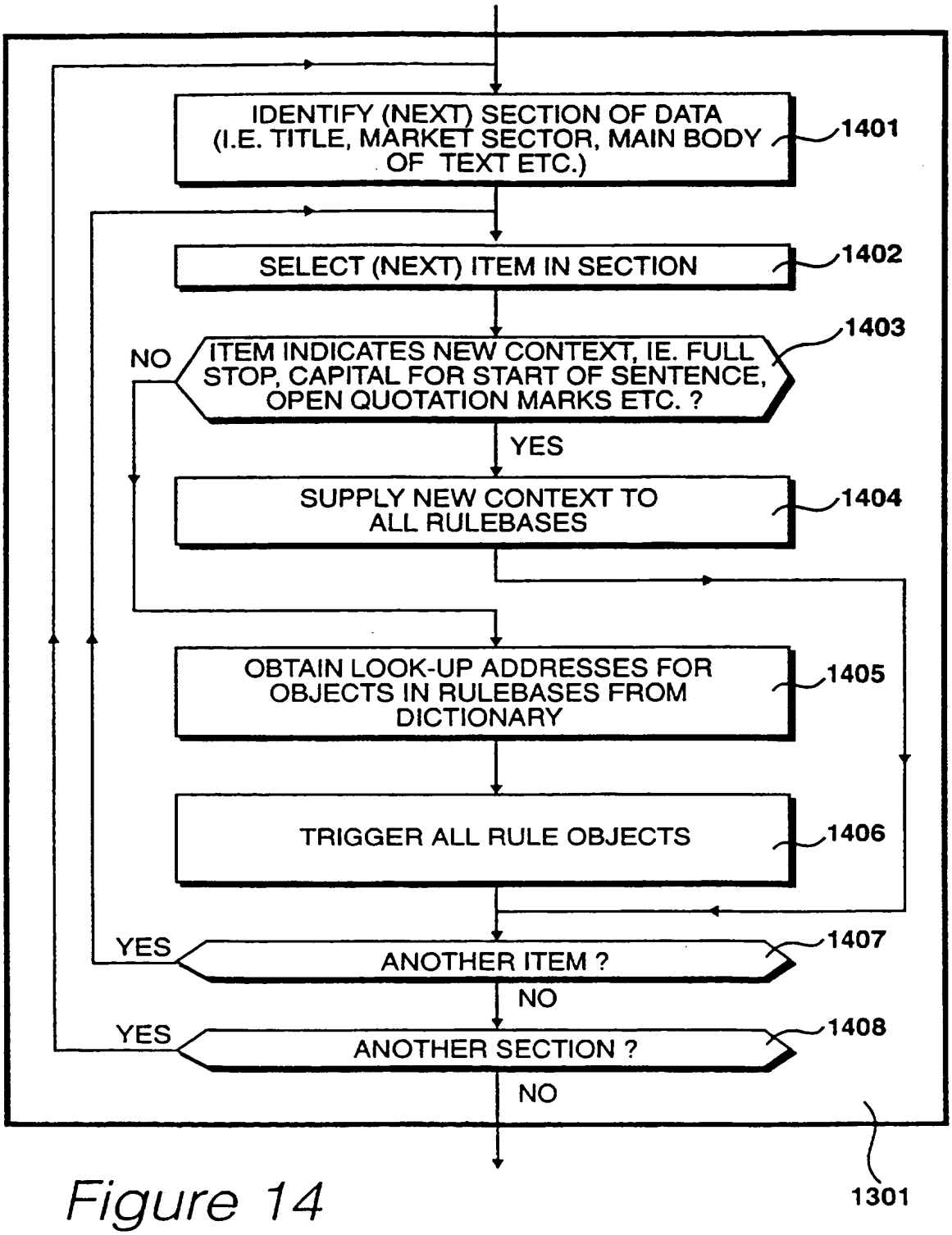
IDENTIFY (NEXT) SECTION OF DATA (I.E. TITLE, MARKET SECTOR, MAIN BODY OF TEXT ETC.)
1401
SELECT (NEXT) ITEM IN SECTION
1402
NO
ITEM INDICATES NEW CONTEXT, IE. FULL STOP, CAPITAL FOR START OF SENTENCE, OPEN QUOTATION MARKS ETC. ?
1403
YES
SUPPLY NEW CONTEXT TO ALL RULEBASES
1404
OBTAIN LOOK-UP ADDRESSES FOR OBJECTS IN RULEBASES FROM DICTIONARY
1405
TRIGGER ALL RULE OBJECTS
1406
YES
ANOTHER ITEM ?
1407
NO
YES
ANOTHER SECTION ?
1408
NO
1301

Figure 14

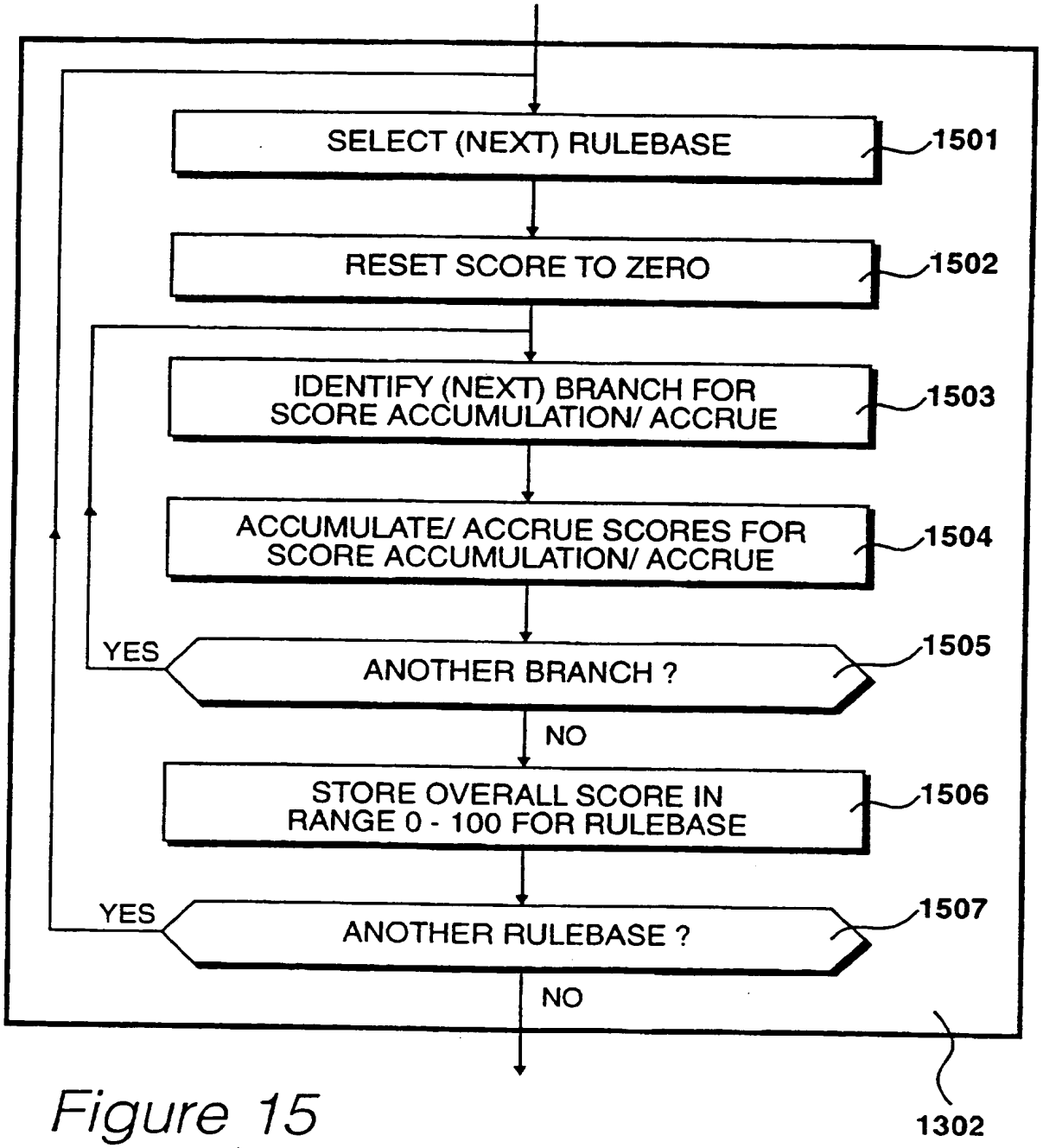
SELECT (NEXT) RULEBASE
1501
RESET SCORE TO ZERO
1502
IDENTIFY (NEXT) BRANCH FOR SCORE ACCUMULATION/ ACCRUE
1503
ACCUMULATE/ ACCRUE SCORES FOR SCORE ACCUMULATION/ ACCRUE
1504
YES
ANOTHER BRANCH ?
1505
NO
STORE OVERALL SCORE IN RANGE 0 - 100 FOR RULEBASE
1506
YES
ANOTHER RULEBASE ?
1507
NO
1302

Figure 15

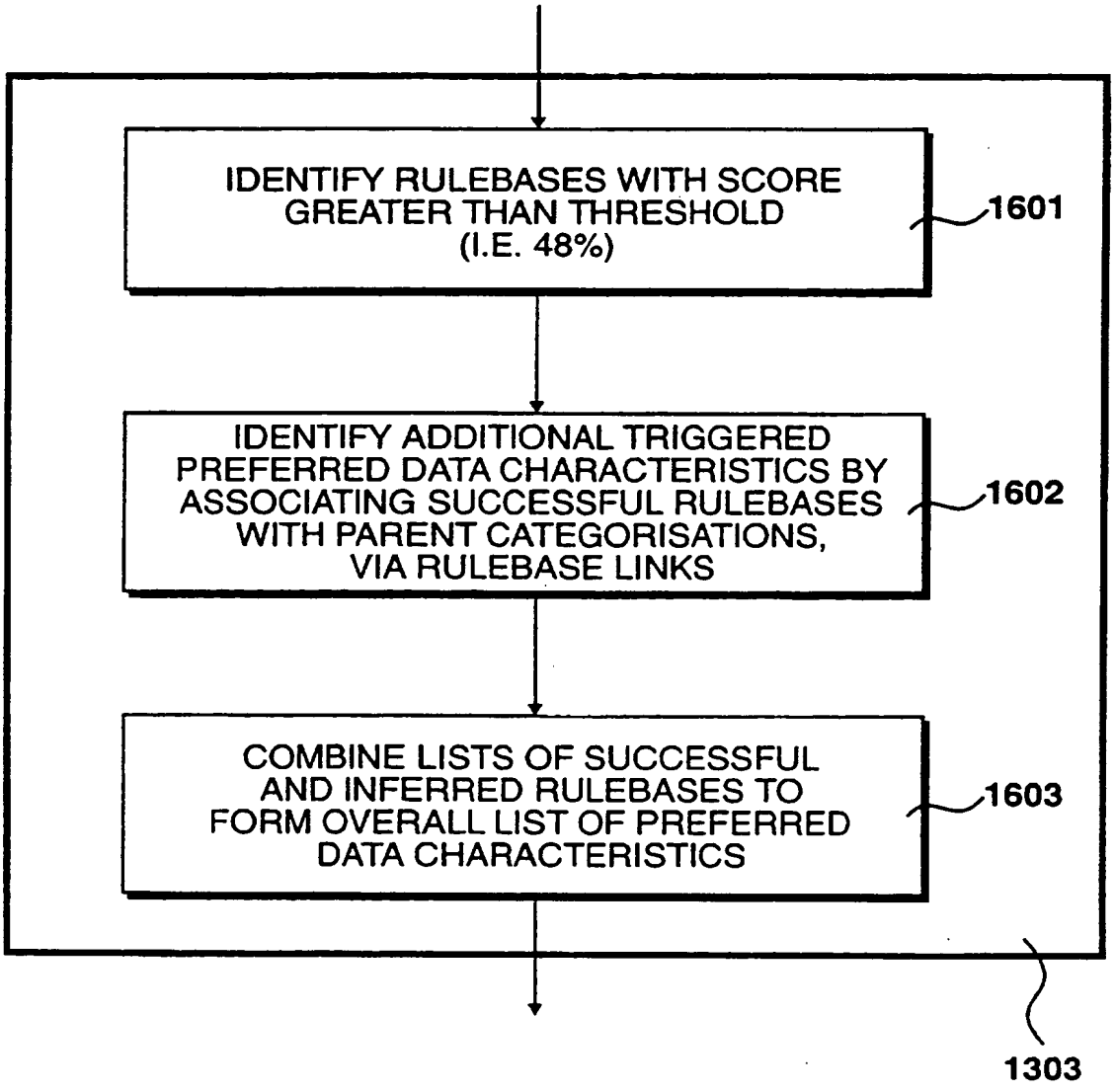
IDENTIFY RULEBASES WITH SCORE GREATER THAN THRESHOLD (I.E. 48%)
1601
IDENTIFY ADDITIONAL TRIGGERED PREFERRED DATA CHARACTERISTICS BY ASSOCIATING SUCCESSFUL RULEBASES WITH PARENT CATEGORISATIONS, VIA RULEBASE LINKS
1602
COMBINE LISTS OF SUCCESSFUL AND INFERRED RULEBASES TO FORM OVERALL LIST OF PREFERRED DATA CHARACTERISTICS
1603
1303

Figure 16

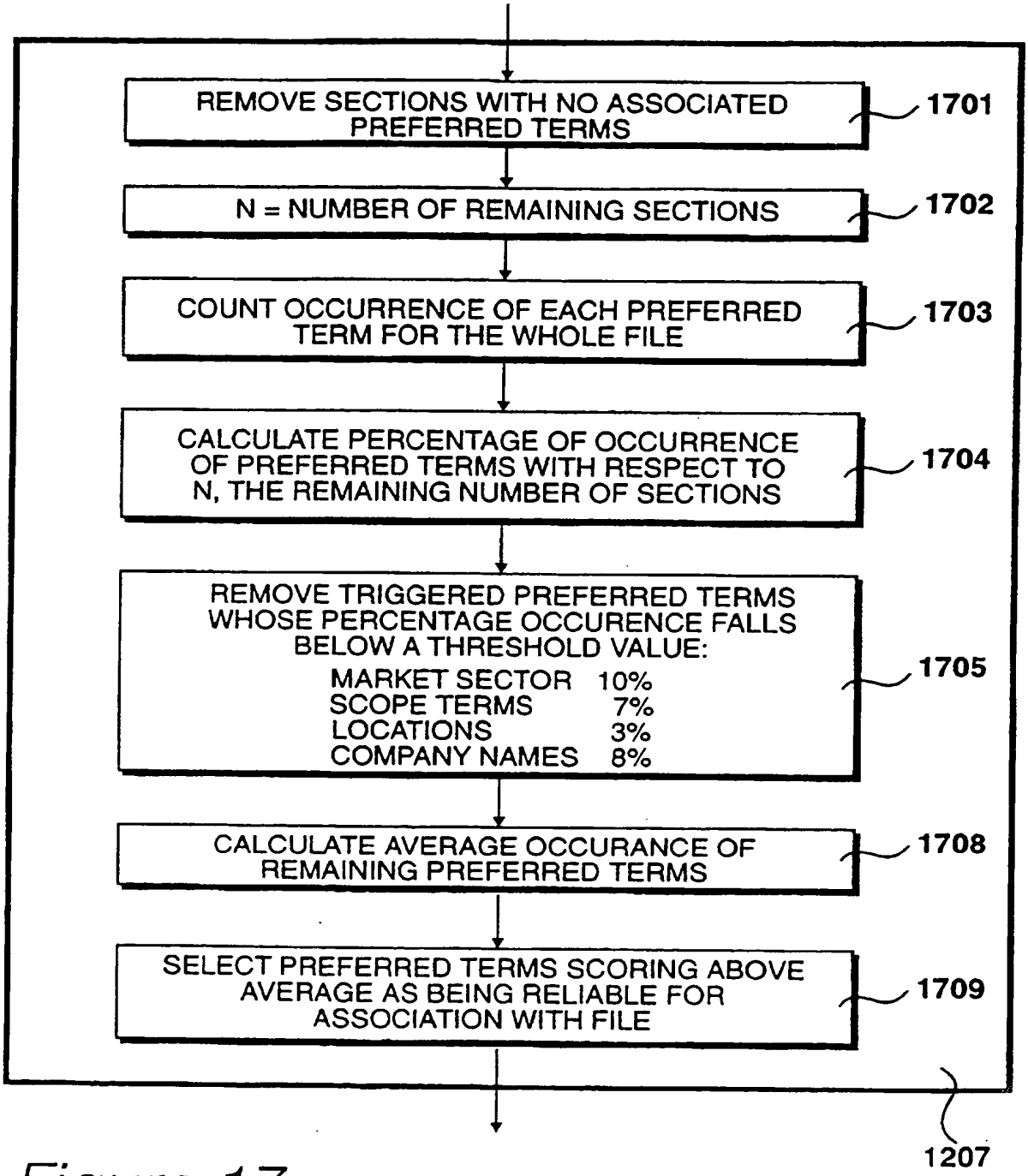
REMOVE SECTIONS WITH NO ASSOCIATED PREFERRED TERMS
1701
N = NUMBER OF REMAINING SECTIONS
1702
COUNT OCCURRENCE OF EACH PREFERRED TERM FOR THE WHOLE FILE
1703
CALCULATE PERCENTAGE OF OCCURRENCE OF PREFERRED TERMS WITH RESPECT TO N, THE REMAINING NUMBER OF SECTIONS
1704
REMOVE TRIGGERED PREFERRED TERMS WHOSE PERCENTAGE OCCURENCE FALLS BELOW A THRESHOLD VALUE:
MARKET SECTOR 10%
SCOPE TERMS 7%
LOCATIONS 3%
COMPANY NAMES 8%
1705
CALCULATE AVERAGE OCCURANCE OF REMAINING PREFERRED TERMS
1708
SELECT PREFERRED TERMS SCORING ABOVE AVERAGE AS BEING RELIABLE FOR ASSOCIATION WITH FILE
1709
1207

Figure 17

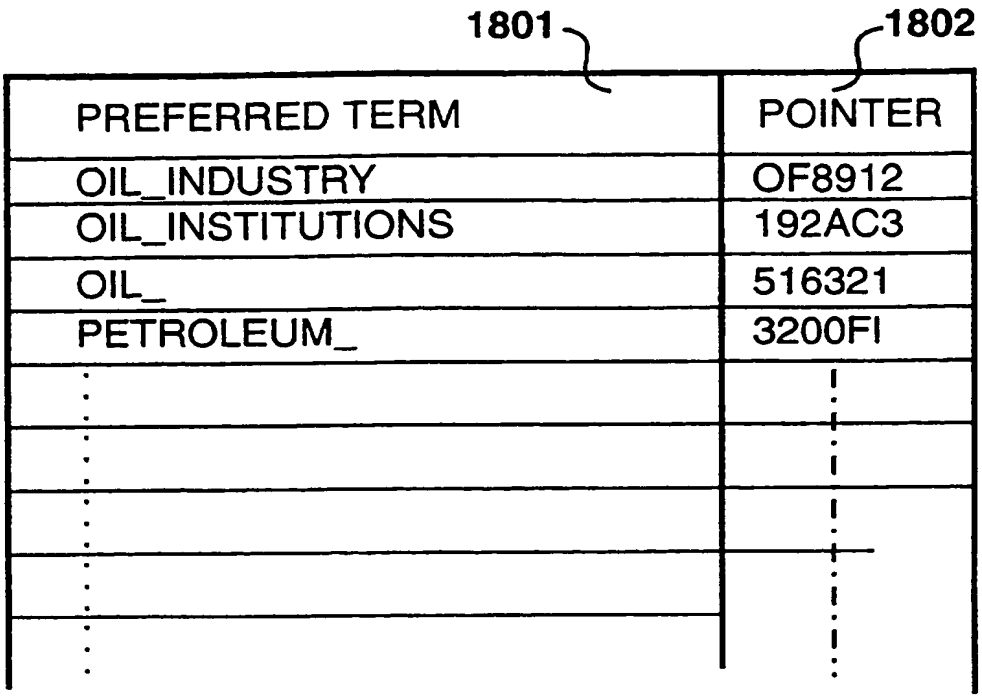

Figure 18

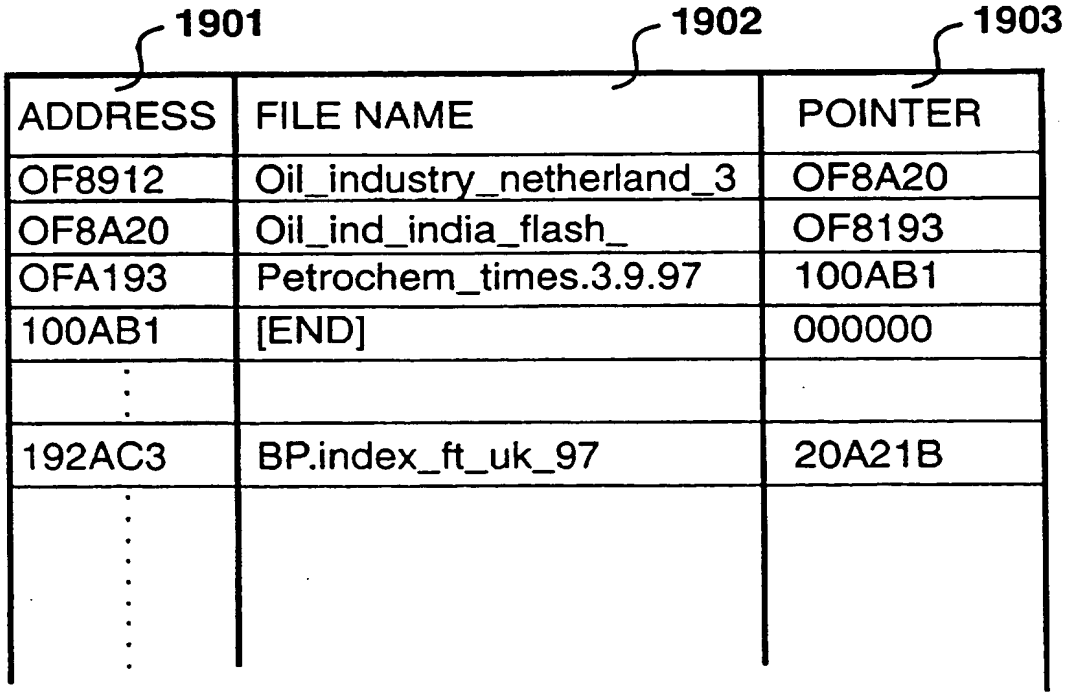

Figure 19

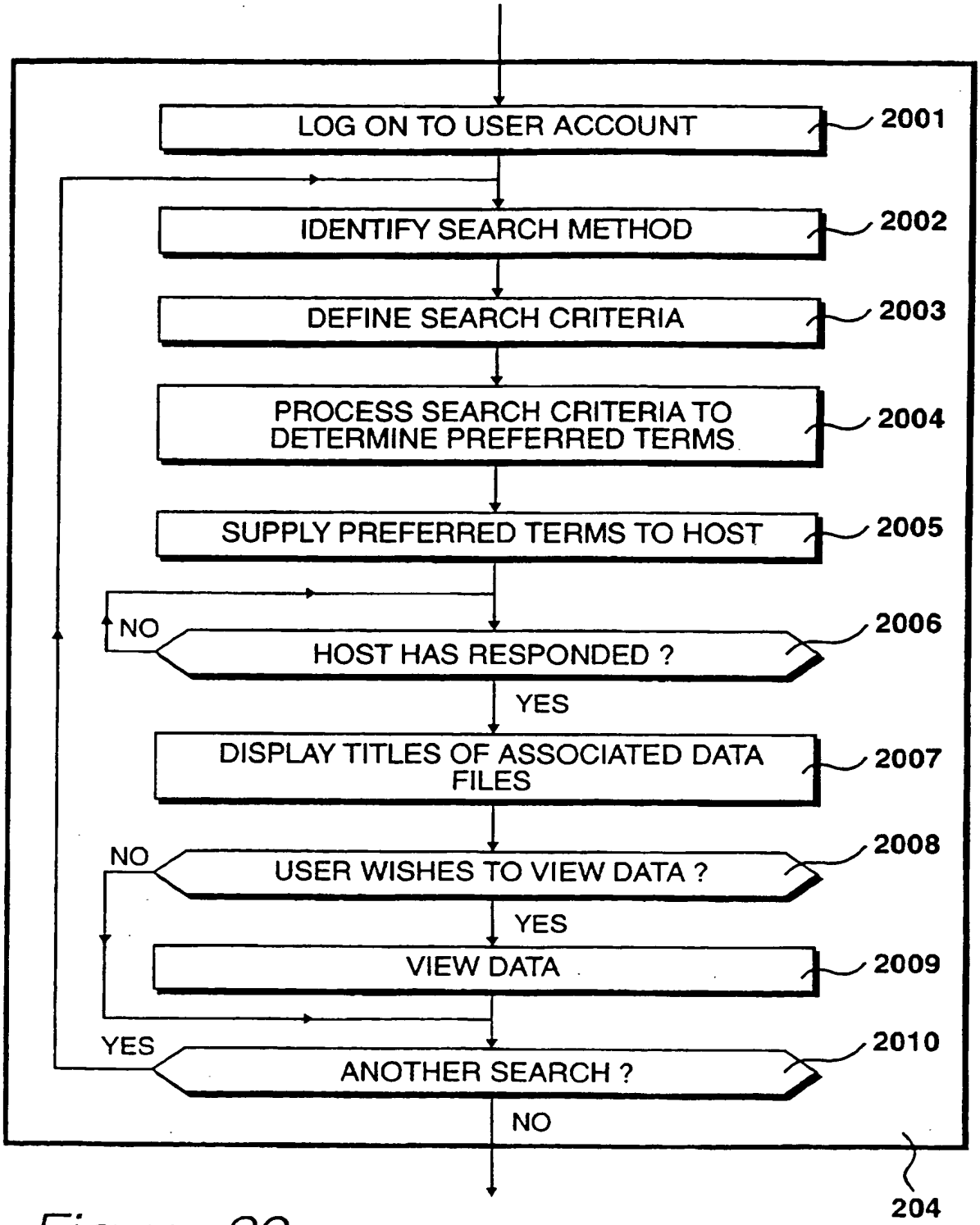
LOG ON TO USER ACCOUNT
2001
IDENTIFY SEARCH METHOD
2002
DEFINE SEARCH CRITERIA
2003
PROCESS SEARCH CRITERIA TO DETERMINE PREFERRED TERMS
2004
SUPPLY PREFERRED TERMS TO HOST
2005
NO
HOST HAS RESPONDED ?
2006
YES
DISPLAY TITLES OF ASSOCIATED DATA FILES
2007
NO
USER WISHES TO VIEW DATA ?
2008
YES
VIEW DATA
2009
YES
ANOTHER SEARCH ?
2010
NO
204

Figure 20

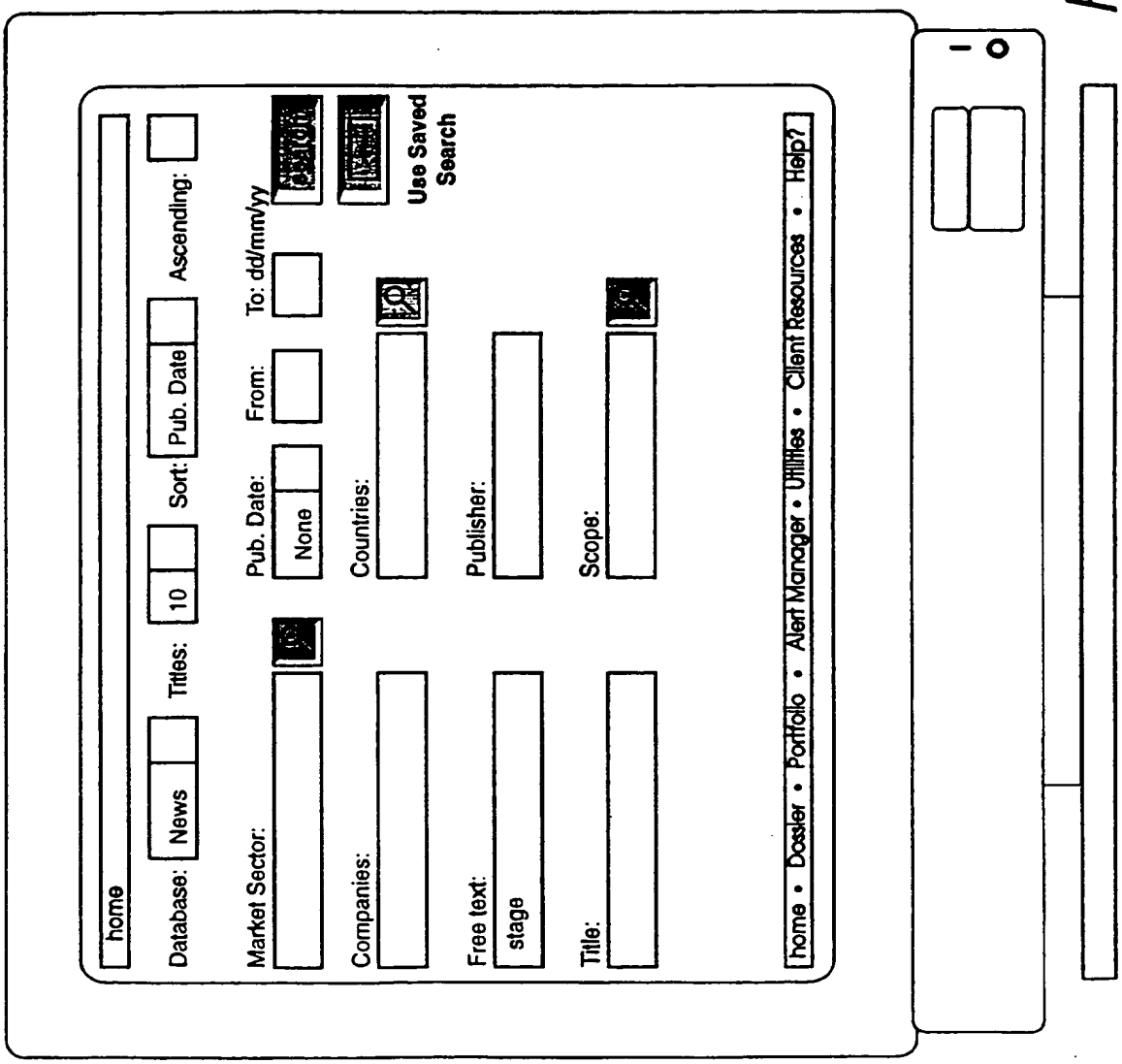
home
Database: News Titles: 10 Sort: Pub. Date Ascending:
Market Sector:
Pub. Date: None From: To: dd/mm/yy
Companies:
Countries:
Use Saved Search
Free text: stage
Publisher:
Title:
Scope:
home • Dossier • Portfolio • Alert Manager • Utilities • Client Resources • Help?

Figure 21

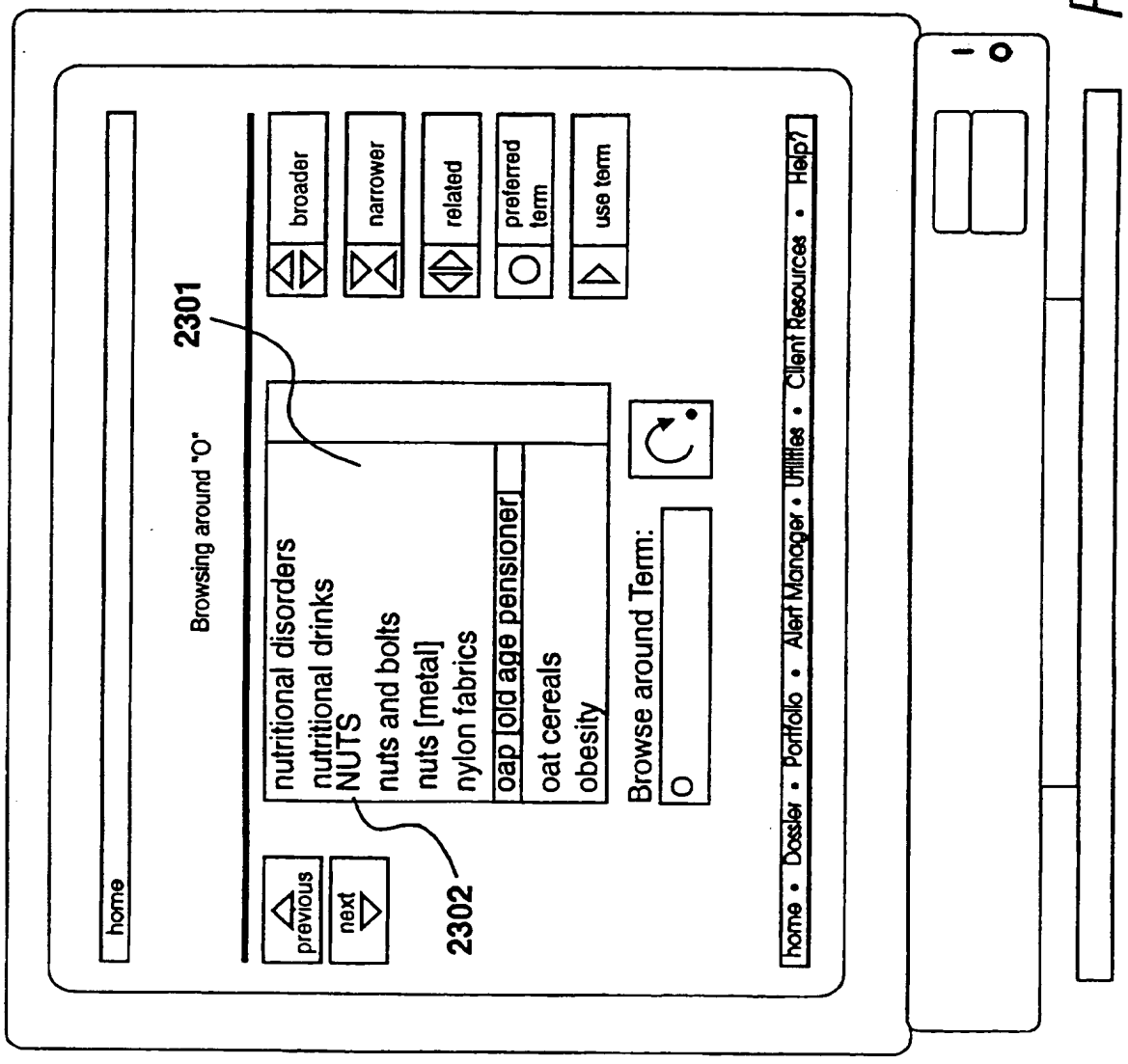
home
previous
next
Browsing around "O"
2301
2302
nutritional disorders
nutritional drinks
NUTS
nuts and bolts
nuts [metal]
nylon fabrics
oap [old age pensioner]
oat cereals
obesity
Browse around Term:
O
broader
narrower
related
preferred term
use term
home • Dossier • Portfolio • Alert Manager • Utilities • Client Resources • Help?

Figure 22

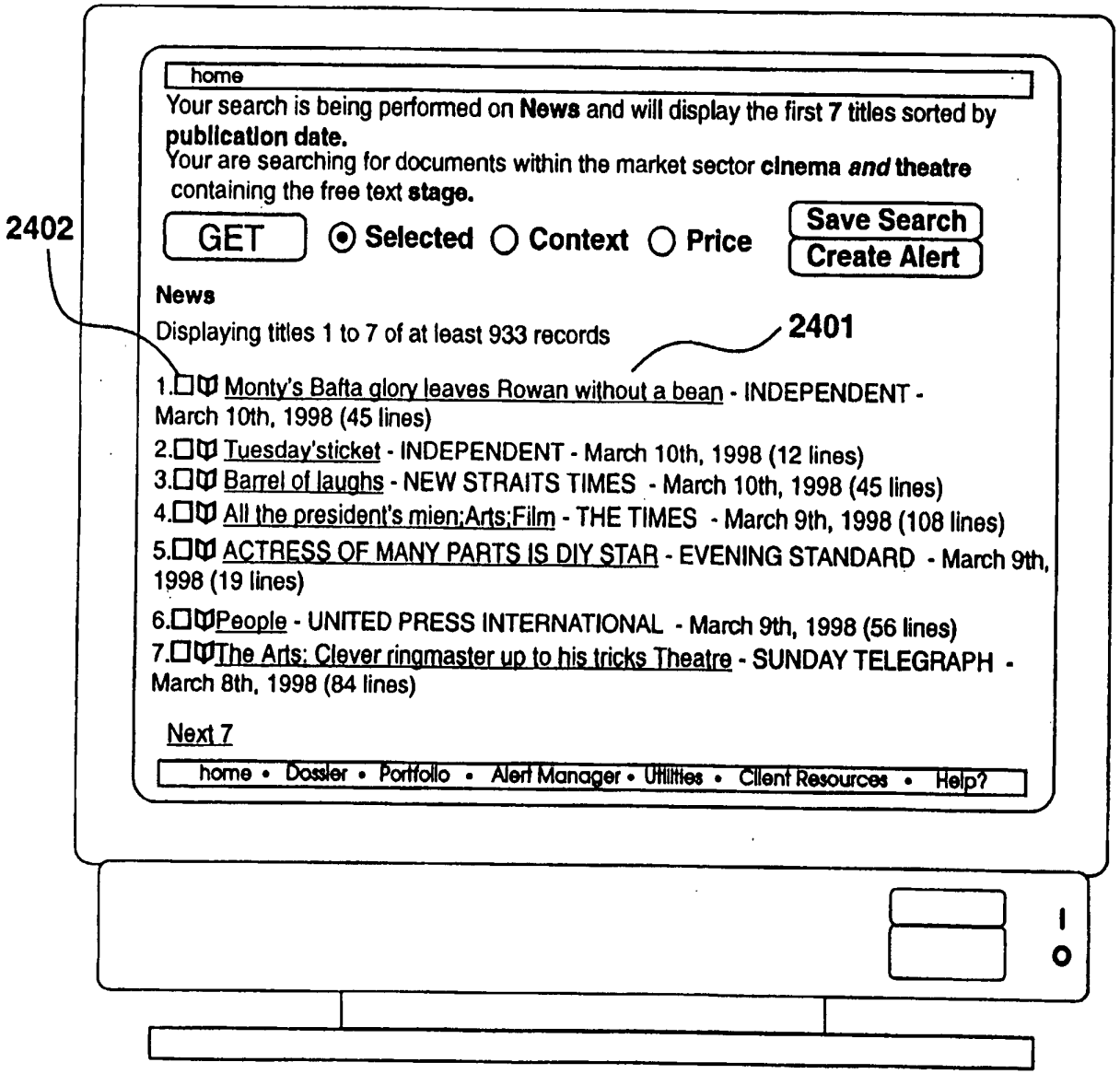
home
Your search is being performed on **News** and will display the first **7** titles sorted by **publication date.**
Your are searching for documents within the market sector **cinema *and* theatre** containing the free text **stage.**
2402
GET
Selected
Context
Price
Save Search
Create Alert
News
Displaying titles 1 to 7 of at least 933 records
2401
1. Monty's Bafta glory leaves Rowan without a bean - INDEPENDENT - March 10th, 1998 (45 lines)
2. Tuesday'sticket - INDEPENDENT - March 10th, 1998 (12 lines)
3. Barrel of laughs - NEW STRAITS TIMES - March 10th, 1998 (45 lines)
4. All the president's mien;Arts;Film - THE TIMES - March 9th, 1998 (108 lines)
5. ACTRESS OF MANY PARTS IS DIY STAR - EVENING STANDARD - March 9th, 1998 (19 lines)
6. People - UNITED PRESS INTERNATIONAL - March 9th, 1998 (56 lines)
7. The Arts: Clever ringmaster up to his tricks Theatre - SUNDAY TELEGRAPH - March 8th, 1998 (84 lines)
Next 7
home • Dossier • Portfolio • Alert Manager • Utilities • Client Resources • Help?

Figure 23

ASSOCIATING FILES OF DATA

Field of the Invention

The present invention relates to associating large data files to information categories.

Introduction to the Invention

Traditionally, database technology has been dedicated to the organisation of numerical and tabular data and it is only recently, particularly with the expansion of the Internet, that demand has grown significantly for the retrieval of text-based files. Several facilities are available on the Internet, commonly referred to as "search engines" which assist in the location of information. The majority of these operate by performing what has become known as "free text" searching, in which a user specifies words which they believe are contained within the target file as a mechanism for instructing the system to retrieve files of interest.

Problems with this technique are well known to users of the available search engines, and a simple enquiry can generate hundreds of thousands of "hits", the majority of which will tend to be totally irrelevant to the user's needs. Furthermore, other relevant files may be missed simply because they do not contain the specific chosen words.

As is well documented, a problem with the Internet is that the freedom of the Internet is also its downfall. Information is not classified before it is made available, therefore it is highly likely that even the simplest search will fail to identify relevant documentation and will take a considerable period of time to perform.

Procedures for classifying volumes of data so as to facilitate subsequent searching are known but these classification processes often involve manual intervention, thereby making them time consuming and prone to human error. Furthermore, except in circumstances where the

documentation is considered to be extremely valuable and will continue to be required over a significant period of time, the cost of performing this manual exercise cannot be justified in terms of the commercial worth of the data sources being considered. Consequently, the problem results in much data being effectively inaccessible and outside the realm of searchable knowledge.

Procedures are known for processing a data file so as to determine whether the data file should be associated with a particular information category. The known processes require a machine readable association file (or outline file) and using this, it is possible for the incoming data file to be processed to produce a numerical score value defining the extent to which the data file is relevant to the associated category. Thereafter, decisions may be made as to whether the data file is to be associated with particular categories by performing respective threshold comparisons.

In practical systems, thousands of such outline files would be required in order to provide a useful level of categorisation. In the present applicant's co-pending British patent application (DGC-P11-GB) a method of generating machine readable association files is described. A plurality of data files are manually selected as being examples of files which should be associated with a particular category. In addition, a plurality of files are selected manually which are considered not to be associated with a particular category. Having identified these files, the process identifies preferred term candidates from the associated files, weights these candidates with reference to files not associated with the category and applies terms to a machine readable association file by analysing the weighting values.

The resulting association files are particularly well suited to associating new data files which are of substantially similar size to the original source data files. Similarly, association files generated by more traditional techniques still tend to be well suited to input data files of a particular size but less well suited to incoming data files of differing sizes. Thus, if a new incoming data

file is larger than the optimum file size, it is possible that many irrelevant files will be inappropriately categorised given that the processing of these files will result in an inappropriately high weighting value being calculated.

Summary of The Invention

According to a first aspect of the present invention, there is provided A method of associating files of data of a size greater than a predetermined size, comprising steps of dividing a file into a plurality of file sections each having a size substantially consistent with a preferred size; categorising each of said file sections to produce sets of section associations; and processing said sets of section associations to produce a set of category associations for the original undivided file.

Preferably, the preferred size is smaller than the predetermined size.

In a preferred embodiment, tables are removed from a data file before a file is divided into sections. Preferably, an assessment is made as to whether it is desirable to increase size sections, whereafter the size of said sections are increased and the dividing process is repeated.

Preferably, data files are continually received from data sources. According to a second aspect of the present invention, there is provided apparatus configured to associate files of data of a size greater than a predetermined size, comprising dividing means configured to divide a file into a plurality of file sections having a size substantially consistent with a preferred size; categorising means configured to categorise each of said file sections to produce sets of section associations; and processing means configured to process said sets of section associations to produce a set of category associations for the original undivided file.

In a preferred embodiment, the categorising means is configured to categorise each file section by processing a section in combination with association files. Preferably, the apparatus includes storage means for storing the association files as outline files and each of the stored outline files

may relate to a respective category.

Brief Description of The Drawings

Figure 1 shows a data distribution environment, including a data processing, storage and retrieval system;

Figure 2A illustrates procedures performed by the processing system shown in *Figure 1*, including a process for specifying preferred terms, a process for associating preferred terms with source files and a process for performing a search in response to a user request;

Figure 2B shows an overview of the operations performed by the environment shown in *Figure 1;*

Figure 3 details a processing system identified in *Figure 1*;

Figure 4 details a process for specifying preferred terms for association with data files identified in *Figure 2*;

Figure 5 details a process for the generation or modification of outline files identified in *Figure 4*;

Figure 6 illustrates a graphical view of an OTL file structure;

Figure 7 illustrates a data file for the information represented graphically in *Figure 6*;

Figure 8 shows a diagrammatic representation of the data file shown in *Figure 7*;

Figure 9 details a subsidiary processor of the type identified in *Figure 3*;

Figure 10 details procedures for the creation of a new OTL file identified in *Figure 5*;

Figure 11 illustrates a plurality of rulebases produced by the execution of procedures identified in *Figure 10*;

Figure 12 details a process for the association of preferred terms identified in *Figure 2*;

Figure 13 details the processing of a section to obtain a list of associated preferred terms identified in *Figure 12*;

Figure 14 details a triggering phase identified in *Figure 13*;

Figure 15 details a scoring phase identified in *Figure 13*;

Figure 16 details a list generation phase identified in *Figure 13*;

Figure 17 details the processing of section results identified in *Figure 12*;

Figure 18 details a table of preferred terms;

Figure 19 details a linked list of preferred terms;

Figure 20 details procedures for performing a search identified in *Figure 2*;

Figure 21 details a screen display prompting a user to identify a search method;

Figure 22 details a screen display prompting a user for search criteria; and

Figure 23 details a screen display for displaying titles of associated files.

Detailed Description of The Preferred Embodiments

The invention will now be described by way of example only with reference to the previously identified drawings.

A data distribution environment is illustrated in *Figure 1* in which data, received from a plurality of data sources **101**, **102**, **103** is supplied to a data processing, storage and retrieval system **104**. Data sources **101** and **102** supply data directly to processing system **104** while data source **103** supplies data via a local area network **105**, thereby allowing user terminals **106** and **107** to gain direct access to their local data source **103**.

The processing system **104** provides access to a plurality of users, such as users **111**, **112**, **113**, **114**, **115**, **116** and **117**. User **111** has direct access to the processing system **104** while users **112**, **113** and **114** gain

access to the processing system **104** via the Internet **118**. Users **115**, **116** and **117** exist within a more sophisticated environment in which they have access, via a local area network **119** to their own local database system **120** in addition to a connection, via an interface **121**, to the data processing system **104**.

All incoming data from data sources **101** to **103** is categorised with a key word in seven separate fields, comprising "market sector", "location", "company name", "publisher", "publication date" and "scope". Users, such as users **112** to **117** may specify almost any term as the basis for a search and are then prompted by an equivalent word or phrase which constitutes more preferred search parameters. For example, a user may specify a search word such as "confectionery" and the system will prompt the user to consider narrower terms such as "chocolate" along with related terms such as "cakes" or "desserts", or broader terms such as "food". From a simple request, a user is given an option of focusing further or of taking a broader overview of the subject under consideration.

The scope of an article refers to the context in which the document or article was written. For example, the scope field may consider questions as to whether the article concerns "mergers and acquisitions" or "seasonal trends" et cetera. Such terms are useful in gathering related information from a wide variety of industries and markets and may prove invaluable for particular applications.

The same criteria used for indexing are offered for search purposes and the same indexing terms are used for all documents across a range of specific databases. An overview of procedures performed by the data processing system **104** is illustrated in *Figure 2*. At step **201** preferred terms for association with data files are specified. This step is essentially performed as an "off-line" process; establishing the environment for allowing source data to be processed as it is received from sources.

Steps **202**, **203** and **204** represent on-line procedures after the preferred terms have been specified at step **201**. At step **202** the processing system **104** receives data from sources such as sources **101**, **102** and **103**. The source data may be transmitted using different protocols, formats and standards therefore the processing system performs a standardisation process so that the data may be stored locally at the data processing system using standardised formats.

At step **203** the data is processed so as to enhance a user's ability to identify information of interest. Files of machine-readable data received from the sources are associated with specific preferred terms which may be considered as defining particular information types. A file is considered and individual data elements, usually in the form of natural language words, are examined to identify occurrences of specified data types. The purpose of this association is to identify files of data which are of interest in relation to particular topics. This enables a user to organise a search which should result in useful information being supplied to said user, with reference to said topics and defined terms, from an extremely large database of stored data files. In this way, the technical procedures performed by association step **203** significantly enhances the overall functionality of the system and provides an industrially applicable approach to allowing highly focused sets of information to be supplied to a user in preference to large volumes of data; much of which will tend to be totally irrelevant.

In order to achieve this, the files of data are processed and are given a score representing a numerical value as to their relevance with respect to the predefined topics. Scores are adjusted in response to the number of identified occurrences of a specified data type. Furthermore, these scores are also adjusted in relation to the size of the data contained within the file. In particular, occurrences of data types in relatively small files are given a higher weighting with occurrences in larger files being given a lower weighting. Thus, the adjustment of scores is related inversely to the actual size of the

data file. Thereafter, a threshold for the scoring values may be set and information types are associated with particular files dependent upon whether particular value scores fall on one side or on the other side of this threshold.

At step **204** a search is performed, in response to preferred terms identified by a user such that information of interest may be identified within the data stored by the data processing system **104** and transmitted to user terminals, such as terminal **111**, over transmission channels as illustrated in *Figure 1*.

An overview of the operations performed by the environment shown in *Figure 1*, in accordance with the present invention, is shown in *Figure 2B*. Processing system **104** receives input files, such as input file **221**. It has been determined that incoming file **221** has a size which is greater than a predetermined size and in this example the predetermined size is set at fifty thousand characters. A process **222** divides the incoming file into a plurality of file sections each having a size substantially consistent with a preferred size. In this example, the preferred size is established at ten thousand characters. Furthermore, a file section is considered to have a size consistent with this preferred size if it has an actual size of between ten and twenty thousand characters.

File sections derived from file **221**, in response to the operation of process **222**, are illustrated generally at **223**. Six file sections are shown but the actual number of file sections produced will depend upon the size of the original file.

File sections **223** are each individually categorised by process **224** and it is assumed that a process is available for performing this categorisation upon files having sections consistent with the preferred size.

Process **224** produces a set of associations for each section and the sets of section associations are processed by process **225** to produce a single set of category associations for the original undivided file **221**. An association process **226** then associates the original file **221** with the set of

associations produced by process **225** so that the file data and its association data may be written to a database **227**.

Processing system **104** is detailed in *Figure 3*. Data signals from data sources **101** to **103** are supplied to input interfaces **301** via data input lines **302**. Similarly, output data signals are supplied to users **111** to **117** via an output interface **303** and output wires **304**. Input interface **301** and output interface **303** communicate with a central processing system **305** based on DEC Alpha integrated circuitry. The central processing system **305** also communicates with other processing systems in a distributed processing architecture. Processing system **104** includes eight Intel chip based processing systems **311** to **318**, each implementing instructions under the control of a conventional operating system such as Windows NT.

An operator communicates with the processing system **104** by means of an operator terminal, having a visual display unit **321** and a manually operable keyboard **322**. Data files received from sources **101** to **103** are written to bulk storage devices **323** in the form of large magnetic disk arrays. Data files are written to disk arrays **323** after these files have been associated with preferred terms, as illustrated at step **203**. These association processes are performed by the subsidiary processors **311** to **318** and the central processing system **305** is mainly concerned with the switching and transferring of data between the interface circuits **301**, **303** and the disk arrays **323**.

The central processing system **305** communicates with the subsidiary processors **311** to **318** via an Ethernet connection **324** and processing requirements are distributed between processors **311** to **318**. Having addressed a subsidiary processor **311** to **318** the transferring of data to an addressed processor is performed. Each individual incoming data file is supplied exclusively to one of the subsidiary processors. The selected subsidiary processor is then responsible for performing the association process, to identify preferred terms relevant to that particular data file.

Thereafter, the associated data file is returned to the central processing system **305**, over connection **324** and the central processing system **305** is then responsible for writing the associated data file to the disk array **323**. In this way, it is possible to scale the degree of processing capacity provided by system **104** in dependence upon the volume of data files to be processed in this way. The central processing system **305** also maintains a table of preferred terms, pointing to particular data files which have been identified as relevant to said preferred terms.

Process **201** for specifying preferred terms for association with data files is detailed in *Figure 4*. At step **401** a preferred term is selected and at step **402** an outline (OTL) file is generated or modified. At step **403** a question is asked as to whether another term is to be processed and when answered in the affirmative control is returned to step **401**, allowing the next term to be processed at step **402**. Eventually, all of the terms will have been processed resulting in appropriate generations or modifications to their related outline files. Consequently, the question asked at step **403** is answered in the negative whereafter at step **404** data structures are initialised by parsing the OTL files generated at step **402**.

Step **402** for the generation or modification of outline files is detailed in *Figure 5*. At step **501** a visual OTL editor is opened resulting in the editor's visual interface being displayed on VDU **321**. At step **502** a question is asked as to whether an existing file is to be loaded for modification and if answered in the negative a new OTL file is created at step **503**. If the question asked at step **502** is answered in the affirmative, step **503** is bypassed and at step **504** modifications or additions are made to the OTL definition. At step **505** the OTL modifications created at step **504** are tested on a sample of test data and at step **506** a question is asked as to whether another modification is to be made. When answered in the affirmative, control is returned to step **504** resulting in further modifications or additions being made to the OTL definitions. When answered in the negative at step **506**, the new or modified

OTL file is saved at step **507**.

When performing modifications or additions at step **504**, a graphical representation of the OTL file data is presented to an operator via the visual display unit **321**. An example of a display of this type is illustrated in *Figure 6*, representing a graphical illustration of a specific OTL file.

The OTL file stores definitions in an hierarchical tree structure and this structure is represented in the graphical view as shown in *Figure 6*. A representation of the tree may be contracted or expanded and the possibility of expanding a particular branch is identified by a plus sign on a particular line, as shown at **601**. Similarly, when a particular branch has been fully expanded, the line is identified by a minus sign as shown at **602**. Definitions within the file consist of rules, words and labels. The labels allow relationships to be defined between various parts of the file and between individual files themselves. The words identify specific words within an input file of interest and the rules define how and what weights are to be attributed to these words. Each rule line includes, at its beginning, a weight value **603** representing the score that will be attributed when a particular rule condition is met. Rules may also have leaves and the rule defines the way in which scores generated from leaves are combined.

OTL file data, represented graphically in the form shown in *Figure 6*, is actually stored in a data file having a format of the type shown in *Figure 7*. The actual data file shown in *Figure 7* corresponds to the data display in *Figure 6* but in *Figure 7* all of the data, some of which has been rolled up in *Figure 6*, is present. The data contained within the file shown in *Figure 7* is manipulated interactively by an operator in response to the graphical interface displayed as illustrated in *Figure 6*. Score values **603** are also identified in the data file shown in *Figure 7*.

Displayed line **601** in *Figure 6* is generated from line **701** of the actual stored data. The syntax of the language used for recording the data, as illustrated in *Figure 7*, may vary and the example shown is specific to this

particular application. However, the underlying functionality of the language may be considered with reference to the diagrammatic representation shown in *Figure 8*.

The outlines analyse data files in order to produce numerical evidence as to the relevance of a particular file with relation to a particular topic. The OTL definitions and structures are determined empirically and would be modified and upgraded over a period of time. The system does more than merely register the existence of a particular word item by placing the word items within an interacting structure; the nature of which is illustrated in *Figure 8*. The particular entry, given label "oil-industry-mkt" relates to marketing aspects of the oil industry and as such can contribute to an overall score as to the pertinence of incoming data to this particular topic. The first line **801** shows that this particular contribution may provide a total score of forty percent. This total of forty percent is then subdivided such that at line **802** the presence of the phase "buying oil from" has a score of fifty percent. Thus, the total contribution made by the presence of this phrase consists of fifty percent of forty percent, i.e. a total of twenty percent being made to the total contribution. Similarly, as shown at line **803** and below, particular words may be identified which result in contributions of sixty percent of thirty percent of forty percent. Thus, a complete OTL file is structured in this way with particular words and phrases making contributions to an overall score value. These words and phrases may also be specified in the rules as making single contributions or being allowed to accrue.

Examples of score value **603** are illustrated in *Figure 8* at **804** to **815**. The hierarchical structure in *Figure 8* consists of a plurality of branches with lowest level entries being considered as leaves. Each leaf has a score value associated with it and values **809** to **815** are leaf score values. Above these, branch score values exist such that score values **807** and **808** exist at the lowest level of branching with score values **805** and **806** being at the next level of branching further up each connected to the highest level of branching

illustrated by score value **804**.

A total score value for a particular occurrence, detected at any level within the hierarchical tree, results in a final score contribution derived from the product of the score value assigned to that particular level with all score levels identified while ascending the tree structure back to its root.

The score contributions are produced and possibly accumulated when occurrences of the specified elements are identified. This provides a numerical weight to assess whether a file being processed should be associated with a particular information type. The present invention, as implemented within the preferred embodiment, further adjusts these score weight in relation to the size of the overall data file. In particular, it is the branch score values (**804** to **808**) which are modified in preference to leaf values (**809** to **815**).

Subsidiary processor **311** is detailed in *Figure 9*. The processor includes an Intel Pentium processing unit **901** connected to sixty-four megabytes of randomly accessible memory **902** via a PCI bus **903**. In addition, a local disk drive **304** and interface circuit **905** is connected to bus **903**. Interface circuit **905** communicates with the TCP/IP network **324**.

Random access memory **902** stores instructions executable by the processing unit **901**, in addition to storing input data files received from the data sources **101** to **103** and intermediate data. Procedures **503** for the creation of a new OTL file are detailed in *Figure 10*.

At step **1001** temporary memory structures are cleared and at step **1002** an OTL description file is selected. At step **1003** an item in the OTL file is identified and at step **1004** a question is asked as to whether the item selected at step **1003** is a rule definition. If this question is answered in the affirmative, a rule object is defined at step **1005**. Alternatively, if the question asked at step **1004** is answered in the negative, to the effect that the item is not a rule definition, a question is asked at step **1006** as to whether the item is a word definition. If this question is answered in the

affirmative, a dictionary link is created at step **1004**.

At step **1008** a question is asked as to whether the item is a label and when answered in the affirmative a new entry is created in a label list, whereafter at step **1010** a question is asked as to whether another item is present. After executing step **1005** or after executing step **1007**, control is directed to step **1010**.

When a question asked at step **1010** is answered in the affirmative, to the effect that another item is present, control is returned to step **1003** and the next item is identified in the OTL file. Eventually, all of the items will have been identified resulting in the question asked at step **1010** being answered in the negative. Thereafter, at step **1011** a question is asked as to whether another OTL file is present and when answered in the affirmative control is returned to step **1002** allowing the next OTL description file to be selected. Thus, this process continues until all of the OTL files have been considered resulting in the question asked at step **1011** being answered in the negative.

For each OTL file considered, by being selected at step **1002**, a rulebase is generated and a plurality of such rulebases is illustrated in *Figure 11*. Thus, a first OTL file processed in accordance with the procedures shown in *Figure 10* results in the generation of a first rulebase **1101**. Similarly, further iterations of the procedures shown in *Figure 7* result in the generation of rulebases **1102** to **1109**. Typically, for a specific installation, in the order of three thousand rulebases would be generated by execution of the procedures illustrated in *Figure 10*.

Rulebases **1101** to **1109** are stored in memory **902**, which also provides storage space for a dictionary **1121**, a label list **1122** and a data buffer **1123**. The dictionary stores a list of words which have importance in any of the stored rulebases. Associated with each word in the dictionary, there is at least one pointer and possibly many pointers, to specific entries in specific rulebases **1101** to **1109**. Thus, the words identified at **803** in *Figure 8* would all be included in dictionary **1121**. Entries within the dictionary **1121**

are implemented upon execution of step **1007** in *Figure 10*. Similarly, execution of step **1009**, creating a new entry in the label list, allows a label to relate to rules that are elsewhere in the tree structure.

Process **203** for the association of preferred terms with source files is detailed in *Figure 12*. At step **1201** a question is asked as to whether the file is larger than a predetermined file size and if answered in the negative, control is directed to step **1209** where the whole file is processed to obtain a list of associated preferred terms.

If the question asked at step **1201** is answered in the affirmative, to the effect that the file is larger than a predetermined size, any tables present within the file are removed at step **1202**.

At step **1203**, a preferred section size is selected and at step **1204** a file section is selected for processing and at step **1205** the section is processed so as to obtain a list of associated preferred terms. At step **1206** a question is asked as to whether another section is present and when answered in the affirmative control is returned to step **1204** where the next section is selected.

Eventually, all of the file sections will have been processed and the question asked at step **1206** will be answered in the negative. At step **1207** section results are processed to select reliable preferred terms. After data association at step **1207**, the data is stored at step **1210** with its associated preferred terms and data pointers associated with the preferred terms are updated at step **1211**.

A plurality of possible techniques are available for dividing large files into a plurality of sections. Firstly, the file could be divided on strict arithmetic grounds with section boundaries being equally spaced and derived purely on a character count. Thus, sections could be divided at two thousand characters or, more consistent with data storage environments, they could be divided at two kilo-bytes; equivalent to one thousand and twenty-four characters.

Alternatively, the file may itself be structured with headings and sub-headings etc. Thus, it may be preferable to divide the file at the start of headings or, if this is not possible, at the start of sub-headings. Experience has shown that the procedures are more effective if the file is divided at positions related to its actual content.

Procedures **1205** for the processing of sections to obtain lists of associated preferred terms are detailed in *Figure 13*. The overall processing is broken down into three major phases, consisting of a triggering phase at **1301**, followed by a scoring phase at **1302** followed finally by a list generation phase at step **1303**.

Triggering phase **1301** is detailed in *Figure 14*. At step **1401** a section of the data, such as its title, market sector or main body of text, is identified and at step **1402** an item of the identified section is selected. At step **1403** a question is asked as to whether the item indicates a new context, which may be considered as a grammatical marker in the form of a full stop, capital, start of a sentence or quotation marks et cetera. When answered in the affirmative new context information is supplied to all rulebases **1101** to **1109** at step **1404** and control is then directed to step **1407**.

If the question asked at step **1403** is answered in the negative, step **1404** is bypassed and a look-up address is obtained for rule objects in rulebases from the dictionary at step **1405**. Thereafter, at step **1406** all addressed objects are triggered and a multiplication of scores is effected by a score weighting factor. Thereafter, at step **1407** a question is asked as to whether another item is present and when answered in the affirmative control is returned to step **1402**.

Eventually all of the items for a selected section will have been considered resulting in the question asked at step **1407** being answered in the negative. Thereafter, at step **1408** a question is asked as to whether another section is to be considered and when answered in the affirmative control is returned to step **1401**. At step **1401** the next section is identified

and steps **1402** to **1408** are repeated. Eventually, all of the sections will have been considered and the question asked at step **1408** will be answered in the negative.

As shown in *Figure 8* each lowest level leaf of the hierarchical tree has a numerical value associated with the identification of a particular item, as identified generally at **803**. If the amount of data contained within a particular file is less than what would generally be accepted, scores are further adjusted in relation to this size so as to improve the association of files with particular information types. This further adjustment is performed at the lowest leaf level (an example of this being level **803** in *Figure 8*) and these leaf values are multiplied by a file weighting factor, derived from the size of the file, which is triggered at step **1406** as shown in *Figure 14*.

Scoring phase **1302** is detailed in *Figure 15*. At step **1501** a rulebase is selected and at step **1502** a score variable is re-set to zero. At step **1503** a branch is identified for score accumulation/accrue and at step **1504** scores are accumulated or accrued from triggered rules attached to the branch. At step **1505** a question is asked as to whether another branch is to be considered and when answered in the affirmative control is returned to step **1503**. A next branch is selected at step **1503** with procedure **1504** being repeated. Eventually all of the branches will have been considered resulting in the question asked at step **1505** being answered in the negative.

At step **1506** an overall score in the range of zero to one hundred is stored for the rulebase and at step **1507** a question is asked as to whether another rulebase is present. When answered in the affirmative control is returned to step **1501** and steps **1501** to **1507** are repeated. Eventually, all of the rulebases will have been considered and the question asked at step **1507** will be answered in the negative.

Phase **1303** for the generation of a list of associated preferred terms is detailed in *Figure 16*. At step **1601** a rulebase is identified having a score greater than a predetermined threshold. Thus, for a particular application a

threshold may be set at forty-eight percent. At step **1602** additional triggered preferred data characteristics are identified by associating successful rulebases with parent categorisations by rulebase links.

At step **1603** lists of successful and inferred rulebases are combined to form overall lists of preferred data characteristics. Step **1603** results in data being generated by a subsidiary processor, such as processor **311**, which is then supplied back to the central processing system **305** over interface **325**.

Process **1207** for the processing of section results to select reliable preferred terms is detailed in *Figure 17*. At step **1701** sections with no associated preferred terms are removed from the scoring process, that is to say they are ignored and at step **1702** a variable N is set equal to the number of remaining sections.

At step **1703** the number of occurrences for each preferred term is counted for the file as a whole, that is to say, individual counts for each set of occurrences are combined. At step **1704** a percentage of occurrences of preferred terms is calculated with respect to N, as calculated at step **1702**. Thereafter, at step **1705** triggered preferred terms are removed if their percentage occurrence falls below a threshold value, defined in terms of the percentage number of times a category should be triggered.

A category could be triggered by each of the sections therefore the total number of possible triggers is equivalent to the number N of remaining sections. Thus, a percentage occurrence value is given by the number of sections which did trigger a particular category divided by the total number of sections then multiplied by one hundred.

The purpose of step **1705** is to remove associations that may be considered as mistakes. Such associations are identified as mistakes if categories are triggered by only relatively few of the individual sections. Process **1701** and process **1705** both remove associations to preferred terms and it is possible that too many associations may be have been removed, a situation that is likely to occur if the section size selected is too small.

At step **1708** the average occurrence of the remaining preferred terms is calculated and at step **1709** preferred terms scoring above the average calculated at step **1708** are selected as being reliable for association with the original large file. As an alternative to rejecting associations falling below the average at step **1709**, such associations may be retained as possibly reliable associations in addition to the reliable associations.

Referring to *Figure 18*, the preferred term "OIL_INDUSTRY" is shown in first column **1801** associated to a pointer 0F8912 in column **1802**. Address 0F8912 is the first in column **1901** of a linked list shown in *Figure 19*. Column **1902** identifies a particular file name and column **1903** identifies the next pointer in the list. Thus, entry 0F8912 points to a particular file with the file name "OIL_INDUSTRY_NETHERLAND_3" with a further pointer to memory location 0F8A20. At memory location 0F8A20 a new file name is provided, illustrated at column **1902** and again a new pointer is present at column **1903**. Eventually, all relevant files will have been considered and the end of the list is identified by address 000000 at the pointer location in column **1903**.

In an active system, the database **323** will be continually updated and users will continually be given access to the database, all under the control of the central processing system **305**. Thus, with reference to *Figure 2*, it should be understood that the association step illustrated at **203** and the searching step at **204** are actually concurrent and will be effected in response to the availability of data and the demand for searching respectively.

Procedures **204** for performing a search in response to a user request are detailed in *Figure 20*. At step **2001** a user logs onto the system and at step **2002** a search method is identified. At step **2003** search criteria are defined and at step **2004** search criteria are processed to determine preferred terms. At step **2005** a list of preferred terms are supplied to the central processing system **305**.

At step **2006** a question is asked as to whether the host has responded and when answered in the affirmative titles of associated data

files are displayed at step **2007**.

At step **2008** a question is asked as to whether the user wishes to view identified data and when answered in the affirmative the data is viewed; after being downloaded over the communication channel, at step **2009**.

At step **2010** a question is asked as to whether another search is to be performed and when answered in the affirmative control is returned to step **2002**.

Step **2002** requires the search method to be identified and in order to achieve this a user is prompted by a screen display of the type shown in *Figure 21*. Thus, a plurality of text boxes are presented to the user inviting the user to specify a search method.

Step **2003** for the defining of search criteria results in the user being prompted by a screen of the type shown in *Figure 22*. Terms providing a basis for the user's search are displayed in a window **2201**. Preferred terms are displayed in uppercase characters, such as the entry shown at position **2202**.

The displaying of titles of associated files at step **2007** results in the user seeing information displayed of the type illustrated in *Figure 23*. Each entry, such as entry **2301**, includes a check box **2302**. Check boxes **2302** allow a particular item to be selected by a user such that the actual information file may be supplied to the user from the central database over a communication channel.

Claims

1. A method of associating files of data of a size greater than a predetermined size, comprising steps of

dividing a file into a plurality of file sections each having a size substantially consistent with a preferred size;

categorising each of said file sections to produce sets of section associations; and

processing said sets of section associations to produce a set of category associations for the original undivided file.

2. A method according to claim 1, wherein said preferred size is smaller than said predetermined size.

3. A method according to claim 1, wherein tables are removed from a data file before said file is divided into sections.

4. A method according to claim 1, wherein an assessment is made as to the desirability to increase size sections, whereafter the size of said section are increased and the dividing process is repeated.

5. A method according to claim 1, wherein data files are continually received from data sources.

6. A method according to claim 1, wherein said categorising is performed by processing a data file in combination with association files.

7. A method according to claim 6, wherein said association files are stored as outline files.

8. A method according to claim 6, wherein each association file relates to a respective category.

9. A method according to claim 1, wherein categories are searched in response to a user request.

10. A method according to claim 9, wherein information identifying files is generated in response to said search and returned to a requesting user.

11. Apparatus configured to associate files of data of a size greater than a predetermined size, comprising

dividing means configured to divide a file into a plurality of file sections having a size substantially consistent with a preferred size;

categorising means configured to categorise each of said file sections to produce sets of section associations; and

processing means, configured to process said sets of section associations to produce a set of category associations for the original undivided file.

12. Apparatus according to claim 11, wherein said dividing means is configured to divide files into file sections each having a size substantially consistent with a preferred size, wherein said preferred size is smaller than said predetermined size.

13. Apparatus according to claim 11, wherein said dividing means is configured to remove tables from a data file before dividing said file into sections.

14. Apparatus according to claim 11, wherein said processing means is configured to determine the desirability to increase size sections, and, if such a determination is made, said dividing means is instructed to increase size sections and repeat the dividing process.

15. Apparatus according to claim 11, including data sources arranged to continually supply data for association.

16. Apparatus according to claim 11, wherein said categorising means is configured to categorise each file section by processing a section in combination with association files.

17. Apparatus according to claim 16, including storage means for storing said association files as outline files.

18. Apparatus according to claim 17, wherein each of said stored outline file relates to a respective category.

19. Apparatus according to claim 11, including searching means configured to search categories in response to a user request.

20. Apparatus according to claim 19, including output means configured to supply output information identifying files selected by said search.

The Patent Office

24

Application No: GB 9808805.7
Claims searched: 1-20

Examiner: K. Sylvan
Date of search: 9 October 1998

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.P): G4A (AUDB)

Int Cl (Ed.6): G06F (17/30)

Other: Online: LISA

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
	None	

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.